عنوان زیرپروژه:

بررسی مستندات ابزارهای خودکار خلاصه‌سازی زبان‌های دنیا برای به کارگیری در خلاصه‌سازی متون زبان فارسی

فهرست مطالب

عنوان **شماره صفحه**

فهرست شکل‌ها

عنوان **شماره صفحه**

1 - مقدمه

یکی از شاخه‌های مرتبط به خلاصه‌سازی متون، متن کاوی است که به معنای کشف اطلاعات جدید به وسیله استخراج اطلاعات به صورت اتوماتیک و بر اساس منابع مختلف می‌باشد. در متن کاوی تعداد زیادی متن تحلیل شده و الگوهای با ارزشی که در متن مخفی است یافته می‌شود. بخش متن کاوی در سال‌های اخیر شامل کشف ارتباط بین کلمات، طبقه بندی اسناد و خلاصه‌سازی شده است که در این تحقیق به مقوله خلاصه‌سازی پرداخته می‌شود.

خلاصه‌سازی عبارت است از نمایش فشرده و دقیق متن ورودی به نحوی که متن خروجی مفاهیم مهم متن ورودی را در برداشته باشد. همگان بر این باور می‌باشند که فرایند خلاصه‌سازی اتوماتیک متون می‌بایست بر پایه فهم کامل متن باشد و فرایندی که انسان در این مورد طی می‌نماید را به نحوی تقلید نماید. به طور کلی خلاصه‌سازی یکی از زیر مجموعه مشکلات پیچیده پردازش زبان طبیعی می‌باشد که در حال حاضر همچنان حل نشده است و هنوز زمان باقی است تا به حدی برسیم که ماشین همانند انسان به طور کامل متن را متوجه شود [1].

هدف خلاصه‌سازی اتوماتیک آن است که مراحلی که توسط فرد انجام می‌شود، در خلاصه‌سازی اتوماتیک انجام شود. بدین صورت که تمام متن خوانده و فهمیده شود و سپس خلاصه تولید شود. فهمیدن متن شامل تشخیص قسمت‌های مهم و غیر مهم متن است. مرحله تبدیل متن ورودی به متن خلاصه شامل مشخص کردن کلمات کلیدی، مفهوم اصلی، کلمه‌های مهم و جمله‌های مهم می‌باشد.

مسائلی که در خلاصه‌سازی می‌بایست در نظر داشت آن است که اولا به چه صورت یک توصیف نرمال از متن ورودی می‌توانیم داشته باشیم. 2) چگونه مهمترین بخش متن را تشخیص دهیم 3) چگونه متن را تجزیه کرده و متن خلاصه را ایجاد نماییم. وجود یک سیستم تطابق دهنده که متن ورودی انتخابی را به پایگاه داده از قبل ایجاد شده مرتبط نماید ضروری به نظر می‌آید.

در این تحقیق ابتدا به مزایا و کاربردهای خلاصه سازی پرداخته می‌شود و پس از آن انواع خلاصه، چالش‌های خلاصه‌سازی و مراحل خلاصه‌سازی بیان می‌شود. سپس به کارهای انجام شده در این زمینه اشاره می‌شود و نهایتا نتیجه گیری از این تحقیق آورده شده است.

2 - مزایای و کاربردهای خلاصه‌سازی اتوماتیک

پیشرفت اینترنت در سال 1990 و پیشرفت روز افزون پایگاه داده‌های الکترونیکی نیاز به توسعه و تکمیل ابزار جستجوی اطلاعات را بیشتر نمود. در انتهای دهه 1990 پروژه‌های تحقیقاتی روی ایجاد سیستم‌های خلاصه‌سازی اتوماتیک در کشور آمریکا و کشورهای اروپایی تعریف شد [2].

از کابردهای خلاصه‌سازی می‌توان به خلاصه‌سازی اتوماتیک اخبار و ارسال آن‌ها از طریق پست الکترونیکی یا پیامک اشاره نمود. از دیگر کاربردهای آن خلاصه‌سازی تحقیقاتی، تجاری و خلاصه‌سازی صفحات وب برای آن‌که در صفحه موبایل قابل نمایش باشد، است.

در یک سیستم بازیابی اطلاعات، نمایش خلاصه‌ای که به صورت اتوماتیک ایجاد شده است مناسب و مفید می‌باشد، در این صورت کاربر می‌تواند سریعا تصمیم بگیرد که چه سندی مطلوب و ارزش باز کردن دارد. گوگل تا حدودی این کار را انجام می‌دهد و اطلاعات مختصری به همراه نتایج جستجو نشان می دهد. به خاطر علاقه روز افزون دولت، بخش بازرگانی، بخش دانشگاهی به این صنعت، تحقیقات و محصولات خلاصه‌سازی زیادی در سراسر دنیا موجود می‌باشد. سیستم‌های سنتی بین سال‌های 1950 تا 1980 برای جستجوی اطلاعات علمی استفاده می‌شد. سیستم‌های خلاصه‌ساز امروزی در فیلدهای جدیدی همانند صنعت مخابرات، ویراستارها، سیستم‌های فیلترینگ و یادگیری زبان خارجی مورد استفاده قرار می‌گیرد. کار اصلی سیستم خلاصه‌ساز کمک به کاربر برای پیدا کردن اطلاعات مورد نیازش است.

3 - عوامل تاثیرگذار در خلاصه‌سازی

محتوای خلاصه بستگی زیادی به ورودی، طبیعت متن، هدف، قصد خواننده و دیگر موارد دارد. عواملی که در خلاصه‌سازی مهم می‌باشد به شرح ذیل می‌باشد [3]:

1) عوامل ورودی: عوامل ورودی به زیر عوامل ذیل تقسیم می‌شود:

- فرم ورودی: شامل طول متن ورودی، ساختار متن (پاراگرافی، نهاد- گذاره‌ای) است. ساختار متن به طور مستقیم روی پردازش متن تاثیر می‌گذارد. زبان متن (انگلیسی، فرانسه و ...) ، نوع متن (متن ادبی،داستانی و ...) در خلاصه‌سازی تاثیر مستقیم دارد.
- موضوع متن: شامل سه مورد عادی، خاص و محدود شده می‌باشد. متن عادی شامل موضوعاتی است که دامنه دانش وسیعی دارند همانند ورزش و باغبانی. متن خاص، متن‌هایی هستند که بستگی به دانش فرد خواننده دارند. متن محدود شده، متنی است که موضوع خاص مربوط به یک سازمان یا انجمن می‌باشد.

2) عوامل خروجی: شامل سه زیر عامل محتوا، فرمت و شیوه است.

- عامل محتوا: این عامل مربوط به آن است که خلاصه عادی یا جواب پاسخ، استخراجی و یا چکیده است. اگر خلاصه عادی باشد تمامی اطلاعات مهم متن را دربربگیرد یا اینکه بعضی از موارد مهم کافی است.
- عامل فرمت: روان، منسجم و یا غیر منسجم بودن خلاصه متن در عملکرد خلاصه‌سازی تاثیر مستقیم دارد.
- عامل شیوه: انواع مختلف خلاصه‌سازی را بیان می‌دارد. شیوه خلاصه‌سازی خبری و یا آگاهی بخش باشد. خلاصه تجمعی، متن خلاصه را از چندین منبع جمع آوری می‌نماید.

3) فاکتورهای هدف: مورد استفاده متن و علت خلاصه‌سازی چه می‌باشد که به سه زیر عامل شنوندگان، شرایط و نحوه استفاده بستگی دارد.

- عامل شنوندگان: دانش شنوندگان در زمینه متن مورد خلاصه تا چه حدی می‌باشد. این مساله به طور مستقیم روی نتیجه خلاصه تاثیر می‌گذارد.

- عامل شرایط: خلاصه قرار است در کجا مورد استفاده قرار بگیرد. اگر زمینه خلاصه در سطع وسیعی شناخته شده باشد جزئیات را حذف می‌نمایند.
- عامل کاربرد: موارد استفاده خلاصه چه می‌باشد. این مساله به طور مستقیم روی نحوه تولید متن خلاصه تاثیر خواهد گذاشت. به عنوان مثال وسیله‌ای برای بازیابی متن است ،جایگزینی برای متن ورودی می‌باشد یا اینکه مروری بر متنی است که قبلا خوانده شده است.

4 - چالش‌های خلاصه‌سازی

4-1- مشکلات خلاصه‌سازی از دید ماشین

4-1-1- پیچیدگی درک زبان طبیعی

برای درک زبان طبیعی به اشکال مختلفی از دانش نیاز است که در ذیل به آن‌ها اشاره می‌شود:

1) دانش صوت شناسی: ارتباط لغات با اصوات را نشان می‌دهد.
2) دانش مورفولوژی: روش ساخت عبارات متنوع و مختلف از روی ریشه کلمات را بیان می‌دارد.
3) دانش نحوی: نحوه ساخت جملات را با استفاده از ترکیب کلمات مختلف نشان می‌دهد. دانش ساختاری هر کلمه موجود در جمله را تعیین می‌نماید.
4) دانش معنایی: معنی هر لغت چه می‌باشد و چطور این معانی ترکیب می‌شوند تا معنی هر جمله تشکیل شود. این دانش، مطالعه معنی جملات را بدون توجه به متن شرح می‌دهد.
5) دانش عملی: جملات چگونه در شرایط مختلف استفاده می‌شوند و هر نوع استفاده از جمله چه تاثیری در تفسیر جمله دارد.
6) دانش سخن: این دانش نشان می‌دهد که معنی و وجود جمله قبلی در تفسیر جملات بعدی چه تاثیری دارد. این اطلاعات خصوصا برای رفع ابهام، تصحیح ضمایر و ارجاعات بسیار مفید است.
7) دانش جهانی: شامل دانش عمومی در مورد ساختارهای اجتماعی، تاثیر عوامل مختلف و موثر بر یکدیگر و دنیایی که کاربران زبان باید از آن برای تفسیر جملات و متون مستقل به زبان استفاده کنند می‌باشد.
8) درک شرایط: معنی جمله مبتنی بر شرایط زمان تولید جمله است. اما معنی یک کلمه به دانشی که در آن زبان، از کلمه وجود دارد برمی‌گردد.

4-1-2- عدم موفقیت در درک متن

مسائل عمده در عدم موفقیت در درک متن به قرار ذیل می‌باشد [3]:

1) عدم موفقیت در تعریف: عدم وجود استاندارد مشخص برای نمایش معنی جمله
2) عدم موفقیت در پیاده سازی توسط قوانین: معمولا برای درک زبان طبیعی از سیستمی استفاده می‌شود که نحوه تجزیه جملات، روش تبدیل به فرم انتزاعی جملات و سایر کارها برای رسیدن به معنا را با قوانین نمایش می‌دهد. نقض قوانین بدلیل مسائل مختلف همانند استثنا و یا برخورد با متون جدید و پیش‌بینی نشده باعث ناتوانی در پیاده سازی این قبیل کارها در زمینه درک متن شده است.
3) نفود و تاثیر زمینه متن: از دیگر دلایل عدم موفقیت، وابستگی معنای واقعی جمله به زمینه سخن است که مثلا از نتایج آن کاربرد یک لغت در متون مختلف است که معنای مختلفی را نسبت به زمینه متن تولید می‌کند. (این مساله بخصوص در زبان فارسی خیلی زیاد مشاهده می‌شود)
4) ابهام: وجود انواع ابهام در متون هم مساله مهم و قابل توجهی می‌باشد.
5) تجزیه درست متن به جملات. تجزیه نحوی نه تنها باید نشانه‌های فرمت گذاری متن و قواعد نقطه گذاری را در نظر بگیرد بلکه کلمه‌های توقف و حروف اختصار را نیز در نظر بگیرد.

از کارهایی که می‌توان در تولید متن انجام داد، ترکیب مناسب جملات و استفاده از شیوه‌های آیین نگارش مانند حذف به قرینه لفظی و معنوی است. تغییر معنای جمله نیز از طریق تعویض قسمت‌هایی از متن باعث بهبود تاثیر بر شنونده خواهد شد.

4-2- چالش‌های زبان فارسی با رویکرد خلاصه‌سازی

برخلاف زبان انگلیسی که در آن هم حروف و هم لغات کاملا متمایز از یکدیگرند، در زبان فارسی پیوستگی میان برخی علائم با لغات وجود دارد و علاوه بر آن تنوع نگارش در کلمات نیز موجود می‌باشد. ریشه یابی فعل که یکی از مراحل مهم پیش پردازش متن برای خلاصه‌سازی می‌باشد، در زبان فارسی

چالش‌های خاص خود را دارد. به عنوان مثال در یک لغت به هم پیوسته هم بن فعل، شناسه، علامت زمان فعل و حتی شناسه‌های مفعولی می‌توان داشت که کار پردازش لغات را پیچیده‌تر می‌نماید به طوری که نمی‌توان از دانش، تجربه و نرم‌افزارهای موجود در این زمینه استفاده نمود و تولید نرم‌افزاری که قادر به حل تمامی این پیچیدگی‌ها باشد، فرایندی زمان‌بر و مستلزم تلاش فراوان می‌باشد. تفاوت‌های ذاتی زبان‌های گسسته‌ای مانند انگلیسی با زبان‌هایی مانند فارسی (عربی و غیره) که با یکدیگر تفاوت‌های بنیادین در قواعد دستوری دارند، منجر به آن شده است که ادعای اعمال تغییرات در ساختار یک نرم‌افزار انگلیسی و به دست آوردن نتایج خوب برای زبان فارسی لزوما امکان‌پذیر نمی‌باشد و مستلزم آزمایش‌های فراوان برای اثبات صحت آن خواهد بود.

4-2-1- چالش‌های نگارش زبان فارسی

مسائل و خصوصیات نگارشی زبان فارسی، امر پیش پردازش اتوماتیک خلاصه‌سازی متون را دچار چالش می‌نماید که از آن جمله می‌توان به موارد ذیل اشاره نمود:

- ساختارهای تولید واژگان در بسیاری از موارد از ترکیب چند واژه‌ی مستقل و با معنی، در کنار یکدیگر به وجود آمده‌اند در حالی که اگر میان این اجزاء فاصله درج شود، تبدیل به دو لغت و اگر نیم‌فاصله درج شود تبدیل به یک لغت خواهند شد. این مسئله در انگلیسی بسیار بسیار کمتر رخ می‌دهد همانند "سیب‌زمینی"
- انواع مختلف نگارش برای یک کلمه همانند اتاق و اطاق و یا به کاربردن همزه به صورت های مختلف همانند مسئول و مسوول
- استفاده از کلمات اروپایی به صورت‌های مختلف همانند سورس و source.
- استفاده از "ا" و "آ" به جای یکدیگر همانند "فرایند" و" فرآیند"
- بازگذاشتن دست نویسندگان در فاصله‌گذاری میان کلمات.
- عدم وجود دستورالعمل قطعی برای استفاده از نیم‌فاصله.
- عدم وجود قواعدی ثابت برای فاصله‌گذاری ترکیبات؛ استفاده از دستورالعمل مبتنی بر لغت (مانند تک‌هجایی بودن، بسیط‌گونه بودن).
- تنوع استفاده از "می" چسبان و غیر چسبان همانند کلمات "می‌تواند" و "میتواند".

- تنوع نحوه به کاربردن "ها" چسبان و غیر چسبان همانند "آنها"، "آن‌ها" و "آن ها".
- تنوع نگارش "ی" اضافه در کلمات مختوم به "ه" همانند "خانه سبز" و "خانه‌ی سبز".

عدم در نظر گرفتن مسائل نگارشی بیان‌شده باعث می‌شود تا تفکیک لغات نا مفید و غیر اصلی از لغات مفید به آسانی انجام پذیر نباشد، ریشه لغات به درستی تشخیص داده نشود و یا لغات همسان، غیر همسان تشخیص داده شود که در مرحله امتیازدهی به لغات تاثیر خواهد گذاشت.

4-2-2- چالش‌های زبان فارسی در بخش ریشه‌یابی

در ریشه‌یابی افعال زبان فارسی که یکی از مراحل اصلی پیش پردازش می‌باشد، چالش‌های ذیل موجود می‌باشد:

1) ابهام در قواعد تولید بن مضارع و بی قاعده بودن برخی افعال. به عنوان مثال ("کردن"، "کرد"، "کن").
2) نیاز به وجود اطلاعات آوایی از بن‌ها برای فرایند تبدیل همانند (کندن، کند، کن)
3) عدم وجود اعراب در زبان فارسی و از آن جایی که یک کلمه بدون اعراب می‌تواند معانی مختلفی داشته باشد، تشخیص فعل از اسم را مشکل می‌نماید.
4) شناسه‌های چسبان به فعل: ابهام میان فعل ماضی ساده و فعل مضارع دوم شخص یا سوم شخص جمع همانند "گردید"، "رنجید"، "گشتید". شناسه‌های چسبان به سادگی قابل تمایز از بخش بن در یک فعل نیستند. زیرا بعضی از آن‌ها در ترکیب با یک بن مضارع منجر به تولید یا بن ماضی می‌شوند.
5) شناسه‌های چسبان به فعل، ابهام میان فعل مضارع با بن منتهی به "ی" در مورد شناسه‌های اول شخص مفرد و جمع همانند گویم، آیم. همچنین افعال مضارع با بن منتهی به "ن" در مورد شناسه‌های سوم شخص مفرد و جمع. و نیز بن‌های ماضی منتهی به (ید، د، ند...). برخی از شناسه‌های چسبان، حروف انتهایی مشترک با هم دارند. در بن مضارع منتهی به "ی" نیز تمایز میان شناسه‌های اول شخص مفرد و جمع نیاز به بررسی دقیق‌تر دارد.

6) حروف التزامی در ابتدای فعل (این حروف در برخی از بن‌ها با حرف آغازین بن یکسان است. در نتیجه بعد از حذف این حروف باید مطمئن شد که باقی مانده یک بن است همانند "بگرد" از "گردیدن" و "ببر" از "بردن".
7) فعل‌های مرکب با پیشوند حرفی همانند "برگرفتن" و "برداشتن"
8) اتصال ضمیر مفعولی به فعل همانند "خورندش" که "آن را خوردند" است. تشخیص اینکه یک فعل به ضمیر مفعولی متصل شده‌است یا خیر برای یک نرم‌افزار واقعا دشوار خواهد بود.
9) مختصر شدن بعضی افعال و اتصال به لغات غیر فعلی همانند "بیدارند"، "خوشحالند". صیغه‌های برخی از افعال مانند "استن" به صورت مختصر نوشته می‌شوند و به لغت قبل از خود می‌چسبند. در نتیجه کشف این افعال بسیار دشوار می‌شود.
10) تعداد زیاد افعال فارسی که کار ریشه‌یابی را دچار مشکل می‌نماید.
11) افعالی با اجزای پخش شده در جمله همانند "من به خانه رفته و غذا خورده‌ام"
12) نبود قواعد ریشه یابی مشخص ابهام در امر پردازش متن را افزایش داده و آن را امری پیچیده نموده است.

4-2-3- چالش‌های ذاتی زبان فارسی

زبان فارسی در ساختار و قاعده (semantic) با زبان انگلیسی متفاوت می‌باشد. برخی از مشکلات ذاتی مربوط به متون فارسی در ذیل طبقه بندی شده است.

1) نبود نکات گرامری تعریف شده همانند آن‌چه که در زبان انگلیسی وجود دارد. این مساله در ریشه‌یابی و پیش پردازش متن تاثیر گذار خواهد بود
2) وجود لغات ترکیبی چند جزئی همانند "آب‌سردکن"
3) در زبان فارسی، ضمائر مفعولی و نشانه‌های جمع به لغات متصل می‌شوند. این مسئله نیز منجر به پیچیدگی تفکیک لغات برای رایانه می‌شود زیرا حروف ضمائر مفعولی و نشانه‌های جمع با تعدادی از وندها و نیز واژگان فارسی مشابهت دارند در نتیجه رایانه نمی‌تواند به سادگی در مورد

بخش اصلی لغت قضاوت کند. به عنوان نمونه، "ان" در "درختان" نشانه‌ی جمع است در حالی که "ان" در "خوران" نشانه‌ی صفت فاعلی می‌باشد.

4) در زبان فارسی قواعد تبدیل بن مضارع به بن ماضی و بالعکس به صورت قانون‌مند وجود ندارد، در نتیجه رایانه برای آنکه بتواند لغات حاصل از بن مضارع یا بن ماضی را تشخیص دهد (مثلاً برای آنکه بداند "ان" در لغت "خوران" نشانه‌ی جمع است یا نشانه‌ی صفت فاعلی)، هیچ راه مشخصی ندارد.

5) ابهام ساختاری به نحوی که یک کلمه می تواند معانی مختلف داشت. بـه عنـوان مثـال، شـیر 3 معنی متفاوت دارد. شیر حیوان، شیر آب، شیر خوراکی.

6) عدم وجود قاعده خاص برای تشخیص اسامی و مکان‌های خاص همانند آن‌چه در زبان انگلیسـی موجود می‌باشد.

7) ابهام در معنی کلمه به علت نبود اطلاعات آوایی همانند"مَرد" و "مُرد"

8) عدم وجود دستورالعمل قطعی برای استفاده از نیم‌فاصله.

9) عدم وجود قواعدی ثابت برای فاصله‌گذاری ترکیبات

10)ناآگاهی رایانه از نقش لغت در جمله همانند "آیین نامه نوشتن" و "آیین‌نامه رانندگی". از آن‌جا که در این ترکیبات، تمام کلمات به تنهایی معنی دارند، رایانه قادر به تعیین مرز لغات نمی‌باشد.

11)وجود کلمه های ترکیبی و امکان در نظر گرفته شدن دو کلمه مجزا. به عنوان مثال سیب زمینی که کلمه مرکبی است که از 2 کلمه سیب و زمینی تشکیل شده‌است.

12)وجود اشتباهات نگارشی در مستندات فارسی همانند نگـارش بـا کاراکترهـای متفـاوت هماننـد آئین‌نامه و آیین‌نامه، به کار بردن همزه به صورت‌های مختلف، و نیاز به یکسان سازی این موارد برای پیش پردازش متن به صورت کارا.

13)ابهام در دستور خط زبان فارسی: پیوسته نویسی کلمات مرکب ترکیب شده با پسوند

14)فقدان هر نوع دادگان بزرگ مانند یک گرامر کامل، یک مجموعه از جملات تجزیه شده نمونـه و یا آمارهای ارزشمند از کاربرد لغات در جملات مختلف و جایگاه‌های مختلف در جمله.

15)همانند سایر زبان‌های هند و اروپایی وجود پیشوند و پسوند و استفاده غلط از واژه‌هایی که طبق عرف تغییر معنا داده‌اند.

16)فقدان یک واژگان کامل زبان فارسی با توجه به شرایط محیطی استفاده از کلمات.

17)فقدان یک ارزیاب اتوماتیک برای هر قسمت از سیستم‌های پردازش زبان طبیعی.

18)فقدان تجزیه کننده اتوماتیک ساختار کلامی متن که قسمت‌های پایه و پیرو متن را تشخیص دهد.

به علت مسائل مطرح شده در فارسی، مرحله پیش پردازش برای امکان پذیر شدن پردازش دقیق متن حائز اهمیت می‌باشد. در مرحله پیش پردازش ابتدا الگوریتم ریشه یابی اعمال می‌شود. اگر همانند انگلیسی ابتدا لیست کلمه‌های توقف اعمال شود، این احتمال وجود خواهد داشت که حروف پیشوندی فعل‌ها به عنوان کلمه توقف در نظر گرفته شوند. پس از آن کلمه‌های ترکیبی تشخیص داده خواهند شد. پس از آن، کلمه‌های توقف در مرحله آخر حذف می‌شوند.

در فارسی به علت نبود یک پایگاه داده لغوی همانند WordNet [4]، اعمال کامل خصوصیات زبان شناختی برای خلاصه‌سازی غیر ممکن می‌باشد. از جمله مسائل دیگر در زبان فارسی نبود حروف کوچک و بزرگ همانند زبان انگلیسی برای تشخیص اسامی خاص (همانند نام روز، هفته، ماه و مناطق جغرافیایی) می‌باشد. تشخیص اختصارات موجود در زبان فارسی از دیگر چالش‌های موجود می‌باشد.

5 - انواع خلاصه

خلاصه به دو نوع استخراجی و چکیده تقسیم می‌شود [5]:

5-1- خلاصه استخراجی

درخلاصه‌سازی از نوع استخراجی از جملات موجود در متن اصلی در خلاصه استفاده می‌شود. مساله ای که در خلاصه‌سازی استخراجی از اهمیت بالایی برخوردار می‌باشد، آن است که فرایند تولید متن خلاصه به درستی انجام نمی شود و از آن جایی که جملات یا پاراگراف‌های کلیدی و مهم آورده می‌شود، خلاصه لزوما پیوستگی مطالب ندارد، زیرا هر جمله از جایی از متن استخراج شده است. . خلاصه‌سازی استخراجی از اطلاعات نحوی استفاده می‌نماید.

خلاصه های استخراجی محدودیت بالایی دارند در حالی که خلاصه هایی که بر اساس تئوری زبان شناختی باشد از کیفیت بالایی برخوردارند [6]. ولی در حال حاضر خلاصه‌های استخراجی به دلیل آسان و ارزان بودن عمومیت بیشتری دارند.

5-2- خلاصه چکیده

در خلاصه‌سازی از نوع چکیده، هدف استخراج مفهوم اصلی متن ورودی می‌باشد بی آن‌که لزوما از جملات متن اصلی استفاده شود. از دو بخش اصلی درک کردن و تولید متن تشکیل شده است. تولید متن شامل ایجاد دوباره محتوای اصلی است که برای آن لازم است تا متن به طور کامل فهمیده شود (پردازش دانش انجام شود) و عناصر بخش‌های مختلف متن با یکدیگر ترکیب شوند به عبارت دیگر پردازش متن تنها بر اساس معنای کلمات انجام نمی‌شود.

انواع خلاصه را از بعدی دیگر می‌توان به دو نوع تقسیم نمود: **خلاصه خبری** و **آگاهی دهنده**. خلاصه خبری بیشتر به این منظور تولید می شود که مشخص شود آیا متن مرتبط به مطلبی که کاربر مد نظرش هست یا خیر. به عبارت دیگر، خلاصه‌های خبردهنده به سر فصل‌های اصلی متن بدون وارد شدن به محتوای آن‌ها می‌پردازد اما خلاصه آگاهی دهنده را به جای متن اصلی عموما می‌خوانند [7].

خلاصه‌های خبری [8] برای خلاصه اسناد علمی غیر وابسته به حوزه مشخص می‌باشند ولی متون روزنامه را نمی‌توانند پردازش نمایند

خلاصه از دیدگاه حوزه می‌تواند مربوط به حوزه خاص باشد (موضوعی) و یا کلی باشد. خلاصه کلی سندها را مستقل از حوزه‌اشان خلاصه می‌نماید، در صورتی که خلاصه موضوعی سندهای مربوط به یک حوزه خاص را شامل می‌شود.

خلاصه‌سازها می‌توانند عمومی و یا موضوعی باشند. تفاوت خلاصه‌سازهای عمومی و موضوعی معمولا به وسیله ساختار کلمه‌های به کار رفته آن می‌باشد. کلمه‌های خلاصه‌سازهای موضوعی، لغت‌های کلیدی برای حوزه خاص هستند در حالیکه لغات خلاصه‌سازهای عمومی، لغات غیر کلیدی و معمول می‌باشند. خلاصه‌سازها می‌توانند متن منسجم یا استخراجی ایجاد نمایند.

6 - مراحل خلاصه سازی

خلاصه‌سازی متن از سه گام اصلی تشکیل شده است. مرحله پیش پردازش متن، مرحله پردازش متن و تولید خلاصه. مرحله پیش پردازش شامل حذف اطلاعات اضافی و غیر مهم و ریشه یابی می‌باشد. به عبارت دیگر، پیش پردازش متن در این بخش انجام می‌شود. بعد از پیش پردازش، بخش مهم متن باقی می‌ماند. بخش اصلی خلاصه‌سازی همان تفسیر متن و در حقیقت مرحله پردازش متن می‌باشد. این مرحله شامل پیدا کردن و امتیازدهی به کلمه‌های مهم می‌باشد. در مرحله آخر، خلاصه متن بر اساس جمله‌ها و امتیازدهی مربوطه ایجاد خواهد شد. مطابق با این مراحل تکنیک پیشنهادی نشان داده خواهد شد.

6-1- پیش پردازش متن

اهمیت تحلیل متن برای هر کاربردی در پردازش زبان طبیعی کاملا روشن است اما اهمیت درک متن برای هر کاربردی متفاوت است. مثلا درک سوال در یک سیستم پرسش و پاسخ برای سیستم خیلی مهم تر از تحلیل و درک متن در یک سیستم تشخیص صحبت می‌باشد. به هر حال مهمترین قسمت در هر سیستم کاربردی پردازش زبان طبیعی تحلیل متن ورودی خواهد بود.سطوح مختلف تحلیل متن که بیانگر هدف تحلیل می‌باشند را می توان به صورت ذیل تقسیم کرد:

- تصحیح متن: تبدیل جمله به فرم استاندارد.
- تجزیه جمله
- استخراج اطلاعات ویژه: هدف در این حالت استخراج اطلاعات خاص می‌باشد
- درک معنی متن: تبدیل هر جمله به یک فرم انتزاعی که بیانگر معنای جمله باشد. هدف از تحلیل در این مورد پی بردن به معنای جمله و یا جملات است.
- یافتن موضوع متن: تفسیر هر قسمت یا بخش از متن برای یافتن مفهوم، موضوع یا عنوانی که متن در توضیح آن آمده است.

- **ü** ارتباط متون: یافتن ارتباط‌های منطقی معنایی و ظاهری بین جملات مختلف بر اساس تشابه یا تفاوت بین جملات یا معنای آن‌ها.

لازم به ذکر است حسب پردازشی که روی متن انجام خواهد پذیرفت ممکن است یک و یا چندین مورد از موارد بالا در مرحله پیش پردازش متن انجام نشود.

بخشی از تحلیل متن استخراج اطلاعات است. این استخراج اطلاعات می‌تواند با استفاده از برچسب زنی کلمات متن باشد. برچسب زنی کمک قابل توجهی به تجزیه صحیح جملات و تصمیم‌گیری در مورد کلماتی که در شرایط مختلف معانی متفاوتی دارند، خواهد نمود. تعیین کلمات کلیدی و استخراج واژه‌های موجود در متن از جمله دیگر کارهایی است که در بخش استخراج اطلاعات در تحلیل متن به کار می‌آید.

در پیش پردازش متن کار یکنواخت‌سازی متن انجام می‌شود. هدف از یکنواخت‌سازی تبدیل متن به یک فرم یکنواخت می‌باشد تا فرایند خلاصه‌سازی بطور صحیح و با دقت لازم انجام شود. از اینرو یکنواخت سازی متن شامل مراحل ذیل می‌باشد:

1) تجزیه مورفولوژی: در تجزیه مورفولوژی هر کلمه از نظر حضور در واژگان بررسی می‌شود. اگر کلمه‌ای در واژگان نباشد از قوانین مورفولوژی برای رسیدن به اصل واژه استفاده می‌شود. مثلا یکی از قوانین این قسمت که برای پیدا کردن لغت اصلی استفاده می‌شود به صورت زیر است:

اسم+ها =>اسم

به عنوان مثال دیگر می‌توان تجزیه افعال مرکب که حاصل صرف فعل می‌باشد را مورد توجه قرار داد. همانند: داشتم می‌رفتم که از مصدر رفتن است.

کار دیگری که در این قسمت انجام می‌شود و از اهمیت خاصی برخوردار است جایگزینی کلمات معادل همانند کلمات unix و یونیکس است. در متون فارسی این اشکال به وفور مشاهده می‌شود و بیشتر به خاطر استفاده از واژه‌های بیگانه و واژه‌های هم معنی که در جامعه جا افتاده نشات می‌گیرد. مساله دیگر آن است که خیلی از واژه‌هایی که از یک واژه ریشه ساخته شده‌اند، معانی متفاوتی دارند. به عنوان مثال می‌توان به حور کلمه داده در "داده‌های کامپیوتری" و "داده شده‌است" اشاره نمود. هر چند که در هر دو حالت، "داده" از ریشه "دادن" استخراج شده است، با اینحال کلمه داده‌های کامپیوتری مبین "اطلاعات داده شده" می‌باشد و لذا یکسان فرض کردن واژه داده در عبارات داده‌های کامپیوتری با "داده" در "داده شده‌است"، صحیح نیست. برای حل این مشکل باید از مدل‌های آماری دو-کلمه‌ای و

قوانین مورفولوژی احتمالی استفاده کرد تا اطلاعات وابسته به متن در اختیار ماشین قرار گیرد و مورفولوژی با توجه به این مدل‌های پیچیده‌تر بطور صحیح صورت بگیرد. مشکل دیگری از این نوع، کلماتی با اشکال متفاوت و معنای متفاوت می‌باشند که از یک ریشه استنتاج می‌شوند. یعنی در اصل کلمات متفاوتی هستند ولی پس از مورفولوژی به کلمات معادلی تبدیل می‌شوند.

2) رفع ابهام: موثرترین انواع رفع ابهاماتی که در خلاصه‌سازی مطرح می باشد رفع ابهام مربوط به ارجاعات و رفع ابهام ضمایر می‌باشد. با بررسی متون فارسی مشخص می‌شود که ضمایر شخصی به ندرت استفاده می‌شوند (بخصوص در کتب) و بیشتر از ضمایر اشاره به همراه نوع اشاره استفاده می‌شود، همانند این سیستم، این تکنیک، همان نتایج. اکثر این ارجاعات رجوع به ماقبل می‌باشد. به عبارت دیگر مناسب‌ترین عبارت مورد رجوع از جملات قبل به نام عبارت کاندیدا انتخاب شود و به جای عبارت مبهم قرار گیرد. تشکیل یک عبارت از جمله ماقبل که شامل کلمه مورد رجوع باشد به اطلاعات دیگری همانند هم‌نشینی کلمات نیاز دارد. مثلا در جملات "سیستم‌های اطلاعات مدیریت ..... ". "این سیستم‌ها....."، تشخیص اینکه سیستم‌های اطلاعات مدیریت مورد ارجاع می‌باشد و نه سیستم‌های اطلاعات با استفاده از چند – وزنی قابل انجام است. بنابراین به طور کلی ابهاماتی که در زمینه تحلیل متن وجود دارد شامل ابهام در تعیین عنوان و موضوع اصلی یک متن می‌باشد، ابهام در معنای جملاتی که چندین معنی متفاوت دارند و همچنین ابهام در تجزیه جمله نیز اتفاق می‌افتد. این مساله در حالتی اتفاق می‌افتد که جمله با چندین قانون قابل اشتقاق باشد و معمولا تنها یکی از این قوانین صحیح است.

3) تشخیص و حذف کلمه‌های بی‌اثر: در این مرحله، کلمات بی اثر که از مرحله قبل برچسب خورده‌اند و یا توسط محاسبات جدید عضو کلاس کلمات بی اثر خواهند شد از متن اصلی حذف می‌شوند.

4) ریشه‌یابی کلمات و یافتن کلمات کلیدی: به دو صورت می‌توان ریشه یابی انجام داد. در نوع اول از الگوریتم‌های ریشه‌یاب استفاده می‌شود و در نوع دوم از پایگاه داده‌هایی که ریشه لغات را داشته باشند استفاده خواهد شد. در زبان انگلیسی از الگوریتم پورتر [9] استفاده می‌شود.

6-2- پردازش متن خلاصه استخراجی

در این بخش به موجودیت‌های متن که در مرحله پیش پردازش متن ایجاد شده‌اند امتیاز داده می‌شود. Mani در سال 1999 سیستم‌های خلاصه‌ساز را بر اساس سطوح مختلف پردازش متن به سه گروه مجزا تفکیک نمود [10] که به قرار ذیل می‌باشد:

6-2-1- روش سطحی

خلاصه‌سازهای سطحی روش‌های آماری، مکانی و تحلیل‌های مکانی- آماری انجام می‌دهند. از یک سری خصوصیات ساده متن برای نمایش اطلاعات محتوا استفاده می‌نماید. این خصوصیات، شامل مواردی همانند خصوصیات کمی متن، فرکانس عبارات، خصوصیات محلی همانند مکان جمله در متن، مکان محلی پاراگراف، خصوصیات زمینه‌ای همانند وجود کلمات عنوان، کلمه‌ها و عبارات خاص در متن می‌باشد.به عنوان مثالی از عبارات خاص می‌توان به خلاصه‌ای که در خود متن موجود است اشاره نمود.

مزیت این نوع روش‌ها آن است که می‌توان روی هر متنی آن را اعمال نمود ولی انسجام و پیوستگی ندارد و ممکن است خوانا نباشد و از روی خلاصه نتوان تعبیر درستی انجام داد.

مشکل اصلی در روش‌های آماری این است که پیوستگی بین مطالب مورد توجه قرار نمی‌گیرند. به عبارت دیگر ارتباط بین کلمات در نظر گرفته نمی‌شوند. به عنوان مثال، ایرادی که به روش محاسبه فرکانس کلمات وارد است آن است که ارتباط بین کلمات را در نظر نمی‌گیرد. و ممکن است این وزن دهی به عبارت برای متن محلی مناسب باشد ولی برای متن سراسری مناسب نیست.

6-2-2- روش موجودیتی- معنایی

تفاوتی که با خط مشی سطحی دارد آن است که یک نمایش داخلی از متن ورودی با مدل کردن متن و ارتباطات بین آن‌ها ایجاد می‌کند. ارتباط بین موجودیت‌ها شامل شباهت، هم معنی، ارتباط نحوی و ارتباط معنایی همانند تضاد، سازگاری متن ، ارتباط لغوی و هم مکانی کلمات می‌باشد.

روش های بر پایه زبان شناسی با تجزیه کلمه و مشخص کردن نحو کلمه آغاز می‌شود و از پایگاه داده لغوی استفاده می‌نماید تا ارتباط بین جملات را استخراج نماید و جملات مهم بر اساس معنایشان انتخاب می‌شوند. خلاصه‌سازهای نحوی – معنایی نحو، معنا و ارتباط نحوی-معنایی در بین اجزای تشکیل دهنده ورودی را در نظر می‌گیرد.

این نوع روش‌ها گران می‌باشند زیرا به منابع دانش پیچیده و تکنیک‌های تفسیر متن پیچیده نیاز دارند ولی عملکرد بهتری نسبت به روش اول دارند. خلاصه‌سازی که بر پایه درک متن باشد نیاز به خط مشی‌های بر پایه دانش و مقایسه بخش‌هایی از متن با پایگاه داده موجود دارد. تحلیل نحوی و معنایی دو بخش مهم این نوع خلاصه‌سازی می‌باشد.

6-2-2-1- روش هم‌مکانی

هم مکانی کلمات بر پایه این اصل است که کلماتی که در یک زمینه متنی مشترک با هم مشاهده می‌شوند، با یکدیگر در ارتباط می‌باشند. چند- وزنی ترتیبی از n کلمه است. چند- وزنی زمانی در یک سند وجود دارد که این کلمات به یک ترتیب پشت سر هم در سند ظاهر شود. بعضی مراجع این توالی کلمات را با در نظر گرفتن ریشه کلمات بدون در نظر گرفتن کلمه‌های توقف محاسبه می‌نمایند.

6-2-2-2- روش‌های مبتنی برگراف

در این روش بعد از گام‌های پیش‌پردازش برای مشخص کردن عناوین (اسامی مهم و).. در متن، سند به شکل گرافی غیر جهتدار که جملات، گره‌های تشکیل دهنده‌ی آن هستند ارائه می‌شود و تئوری گراف می‌تواند به راحتی برای تجسم شباهت بین سندی و درون سندی به‌کار رود [11].

6-2-2-3- روش‌های زنجیره لغوی

زنجیره‌های لغوی به عنوان خوشه‌هایی ازکلمات از نظر معنایی مرتبط با هم تعریف شده‌اند . مثلاً {خانه , سرا ,اتاق }یک زنجیره‌است در حالی که خانه و سرا هم‌معنی هستند، اتاق بخشی از یک خانه است.در

حالت کلی بیشتر خلاصه‌سازهای بر اساس زنجیره لغوی روش یکسانی را در مرحله پردازش متن به صورت ذیل دنبال می‌کنند:

1 -تولید زنجیره‌های لغوی

2-امتیازدادن به زنجیره‌ها

3 -یافتن قویترین این زنجیره‌ها که برای ارزش‌دهی و استخراج

6-2-3- سطح کلامی

روش‌هایی از ساختار کلامی (discourse structure) برای تشخیص بخش‌های مهم متن استفاده می‌نمایند این سطح پردازش متن، ساختار عمومی متن و ارتباط آن‌ها را مدل می‌نماید. این ساختار، شامل هدف سند، موضوعاتی که در متن آمده است و همینطور شامل ساختار کلامی متن می‌باشد.. این مساله که چگونه بتوان یک ساختار را به صورت اتوماتیک استخراج نمود و چگونگی استفاده از آن در تبدیل نحوی اطلاعات مساله مهمی است.

تئوری ساختار کلامی Mann & Thompson در سال 1988 بیان شد. ارتباط موجود بین دو بخش غیر هم‌پوشان متن به نام بخش پایه (nucleus) و بخش پیرو (satellite) بیان می‌شود [12]. تفاوت بین پایه و پیرو در این است که بخش پایه هدف اصلی نویسنده را بیان می‌دارد. مساله دیگر آن است که بخش پایه و بخش پیرو از نظر معنایی و فهمی از یکدیگر مستقل می‌باشند. تئوری ساختار کلامی شامل ساختن درخت دودویی است که ارتباط انسجامی که در متن موجود است را بیان می‌دارد. ارتباط بین واحدهای بیانی تشخیص و برچسب گذاری می‌شود. این واحدها ابتدایی است و برگ‌های درخت نهایی می‌باشد. دو هدف برای تحلیل ساختار کلامی وجود دارد. اول آن است که بخش غیر مهم جمله ها حذف شود و دوم آن‌که ارتباط کلامی جمله‌های باقیمانده حفظ شود تا خلاصه منسجم ایجاد شود. سه سطح ارتباط کلامی پاراگراف، جمله و بخشی از جمله وجود دارد. ارتباط کلامی بین پاراگراف‌های همسایه به این صورت مشخص می‌شود که آیا آن‌ها یک موضوع را بر اساس ضریب هم مکانی‌اشان دنبال می‌نمایند. ارتباط کلامی بین جمله‌ها بر اساس استفاده از هم مکانی معنایی می‌باشد [13].

بر اساس تئوری ساختار زبانی یک متن منسجم می‌تواند به صورت یک درخت ساخت‌یافته در نظر گرفته شود که گره‌های میانی آن ارتباطات کلامی و برگ‌ها گزاره‌هایی باشند که بوسیله بخش‌های موجود در

متن توضیح داده می‌شوند. . تجزیه کننده‌های تئوری ساختار زبانی وجود دارند که درخت تئوری ساختار زبانی مناسبی برای متن ایجاد می‌نمایند. لازم به ذکر است که پارسرهایی در زبان‌های انگلیسی [14]، ژاپنی [15] و پرتغالی[16] وجود دارند که می‌توانند درخت ساختار تئوری زبانی مناسبی ایجاد نمایند.

مطابق با قاعده تئوری ساختار زبانی، تمامی واحدهای گزاره‌ای که در متن موجود می‌باشند می‌بایست به وسیله ارتباط کلامی-بیانی به یکدیگر متصل باشند تا متن پیوسته باشد. اتصال واحدهای گزاره‌ای متن ساختار کلامی/بیانی (rhetorical) را ایجاد می‌نماید. تئوری ساختار زبانی معمولا به وسیله درخت نشان داده می‌شود که هر ارتباطی در آن برای متصل کردن زیر درخت‌ها به کار می‌رود که خود می‌توانند درخت دیگری باشند. در تئوری ساختار کلامی، از این مساله استفاده می‌نماید که نهاد و گزاره در هر جمله یا بخشی از متن موجود است که گزاره قسمت تکمیلی می‌باشد و می‌تواند وجود نداشته باشد.

از یک سری نشانه‌ها همانند "اگرچه"، "بنابراین" و "درنتیجه" برای پیدا کردن ارتباط بین بخش‌های مختلف جمله استفاده می‌نماید ابتدا از یک تجزیه کننده استفاده می‌شود تا جمله بخش بندی و تجزیه شود و از یک معیار پیوستگی همانند همپوشانی کلمات استفاده می‌نماید تا مشخص شود چگونه زیر درخت‌ها به یکدیگر متصل شوند.

25 ارتباط زبانی در تئوری ساختار زبانی وجود دارد که از آن جمله می‌توان به نتیجه، تصدیق اشاره نمود.

تمام روش‌هایی که بر اساس تئوری ساختار زبانی باشند از روش سطحی بهتر هستند. این تئوری به میزان زیادی در محاسبات زبان شناختی استفاده می‌شود و دیگر کارهای انجام شده در حوزه پردازش زبان همانند تفکیک ضمایر و امتیازدهی به سندها را تحت تاثیر قرار داده است [12].

از معایب روش‌هایی که بر اساس تحلیل ساختار متن باشد می‌توان به مشکل بودن آن اشاره کرد. مساله دیگر آن است که این روش، مستلزم پردازش گران زبان‌شناختی می‌باشد. در این نوع خلاصه‌ها ممکن است در مواردی نتیجه دلخواه ایجاد نشود [17].

6-2-4- روش ترکیبی

این روش‌ها از ترکیب 2 یا چند روش خلاصه‌سازی حاصل می‌شود. به عنوان مثال، از ترکیب روش‌های سطحی و روش‌های موجودیتی استفاده می‌شود.

دسته‌ای از خلاصه‌سازها از سیستم‌های یادگیری نیز بهره می‌برند. سیستم‌های یادگیری با یک ساختار ساده شروع می‌شوند و سعی در بدست آوردن دانش بیشتر هستند. به بیان دیگر در پروسه تحلیل متن به صورت تکرار شده تجاربی بدست خواهد آمد. تکنیک‌های ماشین یادگیری در متون ساده کاربرد دارند. یادگیری نظارتی از طبقه بندی برای خلاصه‌سازی بهره می‌برد. این روش به یک نحوی وابسته به حوزه خاص خواهد بود. این روش‌ها به آسانی قابلیت پورت کردن ندارند و پروسه استخراج متن را به نحوی برای حوزه خاص کوچک و خاص می‌نمایند [18-20].

6-3- پردازش متن خلاصه چکیده

در بخش پردازش متن خلاصه‌های چکیده علمی معمولا لیستی از مولفه‌های اطلاعاتی همانند فرضیات، موضوع، متدولوژی، نتایج و یافته‌ها جستجو می‌شود. برای این‌کار به دنبال کلمه‌های خاص در سند می‌گردد تا بتواند ساختار آن را پیش‌بینی نماید.

6-4- تولید خلاصه

در این بخش کارهای ذیل انجام می‌شود:

1) مشخص نمودن اندازه خلاصه: در سیستم‌های خلاصه‌ساز اندازه خلاصه بر اساس تعداد جملات، تعداد کلمات و یا درصدی از سند ابتدایی مشخص می‌شوند. به عنوان مثال خلاصه‌ساز Copernic به طور پیش فرض خلاصه 25% ایجاد می‌نماید و کاربر می‌تواند اندازه خلاصه را بین 5%، 10% و 25% یا 100، 250، 100 کلمه تغییر دهد. این مقادیر نسبی می‌باشند.
2) استخراج جملات بر اساس امتیازهای داده شده
3) ایجاد خلاصه: برای آن‌که پیوستگی متن حفظ شود، جمله‌های استخراجی بر اساس ترتیب‌اشان در متن در خلاصه قرار داده می شوند.

7 - کارهای انجام شده

7-1- خلاصه چکیده

از خلاصه‌سازهایی که از نوع چکیده می‌باشد می‌توان از ProSum که محصول شرکت biritish telecom است نام برد که خلاصه‌سازهای خوبی نمی‌باشند.

7-1-1- خلاصه تجربی

یک نوع از خلاصه‌های چکیده، خلاصه تجربی می باشد. این نوع خلاصه بر این اساس است که خلاصه هر موضوعی می‌بایست شامل موارد کلی باشد. مثلا در مورد زلزله، شدت، مکان، زمان و میزان آسیب رسانی آن مهم خواهد بود. یکی از این سیستم‌ها FRUMP [21] می‌باشد که بر اساس کلمه کلیدی که در متن وجود دارد، ممکن است یک و یا چندین اسکریپت فعال شود. اسکریپت‌های فعال شده کمک به تفسیر متن می‌نماید. اگر بیش از یک اسکریپت فعال شود بر اساس قوانین تجربی یکی از اسکریپت‌ها انتخاب می‌شود. به عبارت دیگر این روش سعی در پیدا کردن اتفاقات مورد انتظار در متن (کلمه کلیدی) می‌گردد و پس از آن متنی که می‌بایست گزارش شود را پیش بینی می‌نماید. 50 اسکریپت وجود دارد که مسلما تعداد زیادی برای خلاصه متون نمی‌باشد.

7-1-2- گرامر موردی

خط مشی گرامر موردی راه خوبی است که به دامنه وابسته نمی‌باشد و می‌توان روی آن پردازش معنایی انجام داد. در گرامر موردی، تعدادی فریم تعریف می‌شود که به یک عنوان و مفهوم خاص مرتبط هستند. بنابراین فریم‌های موردی یا الگوها یک مفهوم را بیان می‌دارند. عناصر جمله در متن ورودی با شناسه‌های موردی مقایسه می‌شوند. در صورتی که با یکدیگر منطبق بودند، از الگوهای موجود استفاده می‌نمایند تا خلاصه تولید شود. با توجه به آن‌چه که بیان شد، از این نوع نحوه تولید خلاصه در یک موضوع محدود که اطلاعات نحوی و معنایی کافی وجود داشته باشد می‌توان استفاده کرد ولی در حالت کلی به علت به

وجود آمدن ابهامات زیاد به این آسانی نیست. یعنی هر چقدر که فریم‌ها بزرگتر می‌شوند انعطاف پذیری‌شان در تطابق دادن به حوزه‌های جدید کمتر می‌شود. به همین علت در خلاصه‌سازی کاربرد زیادی ندارد.

7-2- خلاصه استخراجی

7-2-1- روش‌های بر پایه پردازش سطحی

7-2-1-1- روش‌های بر پایه کلمات و مکان

قدیمی‌ترین کاری که در مورد خلاصه‌سازی انجام شده است به سال 1958 بر می‌گردد [22, 23] که روش استخراج متن خلاصه بر اساس تعداد تکرار کلمه و در حقیقت بر اساس روش آماری بوده است. لهن کلماتی که تا 8 تطابق نیز داشته باشند را یکی در نظر می‌گیرد. فرکانس کلمات محاسبه می‌شود. کلمات با فرکانس کم در نظر گرفته نمی‌شوند. بعد از آن محاسبه می‌شود هر جمله چه تعداد از این کلمات را دارد. روش‌های دیگری [24] [22] [25] بر روی مکان جمله در سند تمرکز نموده‌اند و کلماتی همانند معرفی، نتیجه‌گیری و روش را مورد توجه قرار داده‌اند. به طوری که جملات حاوی این کلمات را در متن خلاصه خواهند آمد. در سال 1969[25] ادمونسون کار لهن را ادامه داد. ادمونسون از ترکیب خطی خصوصیات عنوان، کلمه‌های خاص، مکان و وجود کلمات کلیدی استفاده نموده است و نشان داده که درنظرگرفتن مکان جملات برای متن‌های خبری مناسب می‌باشد.

علاوه بر آن در سال‌های 1970 تا 1980 هم کارهایی در این زمینه انجام شده است ولی کارهای کامل‌تر از سال 1998 شروع شد [26].

- سیستم AutoSummerized که در Microsoft word قرار داده شده است به صورت ذیل خلاصه ایجاد می‌نماید:

1) ضریب وزنی به کلمه‌ها بر اساس تعداد تکرار کلمات داده می‌شود. در شمارش تعداد تکرار کلمات، ریشه یابی و کلمه‌های کمکی در نظر گرفته می‌شود ولی کلمه‌های توقف حذف نمی‌شوند.
2) تجزیه و تقسیم متن به جملات
3) انتصاب ضرایب وزنی به جملات. ضریب وزنی جمله به وسیله جمع ضرایب وزنی کلمات تشکیل دهنده آن بدست می‌آید.
4) امتیاز دهی به جملات بر اساس وزن اشان
5) استخراج N جمله بر اساس اندازه خلاصه. اندازه خلاصه به طور پیش فرض 25% می‌باشد.

-GistSumm [27] خلاصه‌ساز استخراجی می‌باشد که از سه بخش تکه سازی متن، امتیاز دهی به جمله و ایجاد خلاصه تشکیل شده است. امتیازدهی به جملات مطابق با روش بر پایه کلمه کلیدی است، به این صورت که فرکانس تکرار کلمات هر جمله جمع می‌شود و بر اساس طول جمله نرمال می‌شود که به آن روش میانگین کلمه کلیدی گفته می‌شود. جمله‌ای که بالاترین امتیاز را داشته باشد جمله‌ای است که به بهترین حالت متن را توصیف خواهد نمود. انتخاب دیگر جمله‌ها بر اساس میزان ارتباط با این جمله و یا از طریق میزان کل ارتباطی که هر جمله با کل محتوای متن دارد بدست می‌آید. میزان ارتباط با جمله اصلی از طریق میزان مشاهده همزمان کلمات در هر جمله و جمله اصلی مشخص خواهد شد که بر این اساس می‌توان از پیوستگی متن خلاصه مطمئن شد. میزان ارتباطی که هر جمله با کل محتوای متن دارد از طریق انتخاب جملاتی است که امتیاز آن‌ها از یک حدی بالاتر باشد. این حد، میانگین امتیاز کل جملات می‌باشد. GistSumm ارزیابی‌های بسیاری را با موفقیت گذرانده است که از آن جمله می‌توان به DUC 2003 که به تازگی نام خود را به کنفرانس تحلیل متن تغییر داده است، اشاره نمود.

- در روش دیگری موارد ذیل به عنوان خصوصیات پردازش متن در نظر گرفته شده است [28]:

1) شباهت به عنوان: از شباهت cosine استفاده شده است.
2) شباهت به کلمه‌های کلیدی: جمله‌هایی که ارتباط و بستگی بیشتری به کلمه‌های کلیدی دارند، مرتبط‌تر هستند و باید در خلاصه انتخاب شوند.
3) بستگی جمله به جمله: برای هر جمله s ، شباهت بین s و تمام جملات محاسبه و با هم جمع شده و نرمال می‌شود.

4) مشاهده کلمه‌های خاص همانند اسم مکان، انسان و زمان میزان اهمیت جمله را بالا می‌برد.

5) مشاهده حروف اضافه: وجود اطلاعات غیر مهم را در متن نشان می‌دهد.

6) مشاهده اطلاعات غیر لازم همانند "زیرا"، "علاوه بر آن که"در ابتدای جملات.

یک درخت دودویی از روی خصوصیات بالا تشکیل می‌شود که هر سطحی مربوط به یک خصوصیت است. به این صورت، هر جمله‌ای در مسیر درست خود در درخت قرار خواهد گرفت. پس از انجام یک سری پیش پردازش همانند حذف کلمه‌های توقف، ریشه یابی و یکسان سازی، برای هر جمله‌ای خصوصیات بالا را مشخص کرده‌اند و سپس از طبقه بندی کننده‌های c4.5 و neive bayse استفاده کرده‌اند و در مجموعه آموزش و تست آن را اعمال کرده‌اند و در مرحله بعدی به جای آن‌که مقادیر این 5 خصوصیت را برای هر جمله صفر یا یک در نظر بگیرد به صورت فازی در نظر گرفته شده است. برای ارزیابی 10 متن از تافل انتخاب شده و به 5 نفر داده شده‌است تا متن را خلاصه نمایند و نتایج با روش برداری و فازی مقایسه شده و نتایج بهتری داشته.

-[29] یک روش خیلی قدیمی و ساده است که امتیازهای بالا به جملات اولیه داده می‌شد که مختص متون خاص همانند متون خبری می‌باشد. نام این روش lead based است. در این روش، خلاصه بر پایه اولین جمله هر پاراگراف تولید می‌شود. این خلاصه‌ساز ابتدا اولین جمله هر پاراگراف را با استفاده از نشانه نقطه که مرزهای جمله را مشخص می‌کند، استخراج می‌نماید. در ادامه جمله‌های مربوط استخراج شده و در خلاصه آورده می‌شود و در صورتی که تعداد جملات خلاصه کمتر از تعداد پاراگراف‌ها باشد، پاراگراف‌های انتهایی در نظر گرفته نمی‌شود و در خلاصه آورده نمی‌شوند.

7-2-1-2- خلاصه‌ساز FarsiSum

یک خلاصه‌ساز فارسی مطرح به نام FarsiSum [30] وجود دارد که بر پایه خصوصیات آماری متن پایه گذاری شده است و خصوصیات زبان شناختی متن را به جز لیست کلمه‌های توقف برای پیش پردازش در نظر نمی‌گیرد.

FarsiSum نسخه تغییر یافته‌ی سیستم خلاصه‌سازی متون سوئدی به نام SweSum [31] برای پوشش زبان فارسی می‌باشد. خلاصه‌ی خروجیSweSum از نوع مستخرج است و برای زبان های سوئدی ، نروژی ، دانمارکی، اسپانیایی ، انگلیسی ، فرانسوی و آلمانی پیاده‌سازی یا نمونه سازی شده است. این خلاصه‌ساز،

متنی در فرمت HTML یا text را به عنوان ورودی از روزنامه یا گزارش دریافت کرده و خلاصه‌ی آن را تولید می‌کند به این منظور ابتدا محدوده‌ی جملات و کلمات را تعیین کرده و کلمات مخفف شده برچسب می خورند. سپس فرکانس ریشه‌ی کلمات دارای مقوله های نحوی کلاس باز(مانند اسامی، افعال و صفات) با استفاده از یک واژگان ایستا محاسبه می شود. پس از آن نوبت به محاسبه‌ی امتیاز جملات متن است . سیستم برای این امتیازدهی موارد زیر را لحاظ می‌کند:

-اولین خط در متن باید در خلاصه وارد شود ، زیرا امتیاز بسیار بالایی دارد (مقدار پیش فرض آن 1000 است).

امتیاز موقعیت به نوع متن بستگی دارد. ،SweSum دو نوع متن روزنامه و گزارش را پشتیبانی می‌کند . موقعیت خط در گزارش روزنامه اهمیت کمتری دارد.در روزنامه مهم ترین بخش متن، اولین خط به اضافه‌ی چند خط دیگر به ترتیب نزولی است.

-مقادیر عددی مثل تاریخ ها در یک سند مهم است. یک مقدار ثابت به امتیاز خط شامل اعداد اضافه می شود (مقدار پیش فرض آن 1 است).

- متن پررنگ در سند HTML مهم است و امتیاز بالاتری دارد. در متن برخی از روزنامه‌های سوئدی آن آغاز یک پاراگراف را نشان می‌دهد و اولین جمله هر پاراگراف معمولاً پاراگراف را معرفی می کند.(مقدار پیش فرض آن 100 است)

-جملات شامل کلمات کلیدی (کلمات با بیشترین تکرار در متن) امتیاز بیشتری از جملاتی که کلمات کلیدی کمتری را شامل می‌شوند، دارند. این کلمات توسط کاربر از طریق سوال به سیستم داده می‌شوند.

-تمام پارامترهای بالا در تابع ترکیبی با وزن‌های قابل اصلاح (مقدار پیش فرض) برای به‌دست آوردن امتیاز کل هر جمله گذاشته می‌شود. در این تابع امتیاز هر کلمه از حاصلضرب تعداد تکرار آن در وزن کلمه‌ی کلیدی حاصل می‌شود و امتیاز جمله از حاصل نرمال‌سازی مجموع امتیاز کلمات آن بدست می‌آید.

-جملات طبق امتیاز محاسبه شده مرتب شده و خطوطی که بیشترین امتیاز را دارند، تا سقف اندازه برش تعیین شده توسط کاربر(درصد کلمه یا حرف در خروجی) در فایل خروجی نوشته می شوند. این فایل شامل تمام خطوط غیر متنیHTML، تمام خطوط متن امتیاز داده شده و اطلاعات آماری در مورد خلاصه، تعداد کلمات، تعداد خطوط، کلمات با بیشترین فرکانس و ... می‌باشد.

در FarsiSum برخی پیمانه‌ها (مانند قطعه بند) برای کار با سندی با محتوایUnicode و رمزگذاری UTF- 8و نگارش راست به چپ اصلاح شده است همچنین FarsiSum به دلیل عدم دستیابی به واژگان فارسی به جای استفاده از لیست کلمات و مخفف ها تنها با تهیه لیست کلمات غیر مفید فارسی (200 کلمه از پرتکرارترین کلمات فارسی شامل افعال، ضمایر، قیدها، حروف ربط، حروف اضافه و حروف تعریف) عمل پردازش را انجام می‌دهد. در ارزیابی این سیستم از روش مقایسه با خلاصه مرجع دستی با تعداد پایین شرکت کننده‌ها و متن‌های با رمزگذاری8 -UTF استفاده شده است. FarsiSum از روش کلاسیک بدون توجه به معنای کلمات استفاده می کند.

[32] به چکیده سازی از چند منبع فارسی ورودی توجه دارد

خلاصه‌سازهایی که توسط نویسندگان آن ادعا می‌شود که وابسته به حوزه خاصی نیستند همانند FociSum [33] تنها متون روزنامه را می‌توانند خلاصه نمایند و متون علمی را نمی‌توانند خلاصه نمایند.

7-2-2- روش‌های موجودیتی- بر پایه معنی

7-2-2-1- استفاده از زنجیره لغوی

این راه حل‌ها خصوصیات زبان شناختی یک زبان خاص و ارتباط معنایی بین کلمات را در نظر می‌گیرند[34-36]. WordNet [4] یک پایگاه داده لغوی شامل بیش از 118000 کلمه انگلیسی می‌باشد و انواع ارتباطات بین کلمات همانند هم معنی، متضاد، ارتباط جز به کل و را در بردارد. در این راه حل‌ها یک زنجیره لغوی بر اساس کلمات موجود در متن ایجاد می‌شود. زنجیره لغوی، رشته ای از کلمات می‌باشد که از طریق یکی ازانواع رابطه‌های بیان شده با یکدیگر در ارتباط می‌باشند. اُستفاده از زنجیره‌های لغوی به کشف ابهام و تشخیص ساختار متن کمک می‌نماید. مساله اصلی در این روش‌ها انتساب کلمات به مناسب ترین زنجیره و الگوریتم امتیازدهی به زنجیره‌های لغوی می‌باشد. Brazily [36] برای اولین بار امکان ساختن زنجیره‌های لغوی را مطرح نمود. پیاده سازی‌های بعدی آن بیشتر روی رفع ابهام کلمات با معانی مختلف متمرکز شد. راه حل‌هایی نیز وجود دارد که بر اساس پایگاه داده لغوی HowNet پایه ریزی شده است. این پایگاه داده شامل کلمات چینی و ارتباط بین آن‌ها و کلمات

انگلیسی متناظر می‌باشد. در[10] گرافی بر اساس ارتباط بین کلمات ایجاد می‌شود. این ارتباط می‌تواند آماری یا مفهومی باشد.

- [37] روش وزن گذاری جدید بر اساس پیوستگی لغوی در متن ارائه نموده است. از زنجیره لغوی استفاده شده است و روش ابهام زدایی زنجیره لغوی جدیدی که وابستگی بین کلمه‌ها و خصوصیات WordNet را در نظر می‌گیرد ارائه شده است. نوآوری که داشته این بوده که از کلمه‌های بیان کننده موضوع استفاده نموده است، زیرا این کلمات ارتباط معنایی بیشتری نسبت به دیگر کلمات دارند. یکی از روش‌های کارا برای محاسبه پیوستگی لغوی استفاده از زنجیره‌های لغوی می‌باشد. زنجیره لغوی گروه یا رشته‌ای از کلمات است که از نظر معنایی با یکدیگر در ارتباط می‌باشند. [38, 39] روش‌های مناسبی برای استفاده از زنجیره لغوی پیشنهاد داده‌اند. در استفاده از زنجیره لغوی دو مشکل را می‌بایست در نظر داشت:

1) ابهام زنجیره لغوی: ارتباط معنایی بین کلمات بستگی به معنای هر کلمه دارد. اگر ابهام معنایی کلمه حل نشود، زنجیره لغوی ابهام دارد.
2) مربوط بودن زنجیره لغوی: تمام زنجیره‌های لغوی به موضوع سند مربوط نمی‌باشند. روشی برای تفکیک زنجیره مرتبط از غیر مرتبط می‌بایست وجود داشته باشد.

از زنجیره لغوی brazily استفاده نموده است[39] ولی رفع ابهام آن به نحو دیگری می‌باشد. از اصل هم‌مکانی کلمات استفاده نموده‌است. ارتباط هم مکانی بین دو کلمه بیانگر آن است که ارتباط معنایی با یکدیگر دارند.

امتیاز ارتباط کلمه را به صورت ذیل تعریف می‌نماید که در آن w_1 و w_2 دو کلمه در WordNet می‌باشند و $p(w_1,w_2)$ احتمال هم مکانی این دو کلمه می‌باشند. Ns(w) فاکتور نرمال سازی می‌باشد.

$$Assoc(w1, w2) = \frac{\log(p(w1,w2)+1)}{N_s(w_1)*N_s(w_2)})$$

عمق در سلسله مراتب WordNet: کلمه‌های موضوعی در سلسله مراتب WordNet در سطوح پایین‌تر قرار می‌گیرند زیرا معنی خاصی دارند و مفهوم کلی ندارند. امتیاز عمقی به این صورت تعریف می‌شود.

$$DeptScore(w_1, w_2) = Depth(w_1)^2 \times Depth(w_2)^2$$

Dept(w) عمق کلمه w در WordNet می‌باشد.

ضریب ارتباط هم معنی بودن را بر اساس تجربه 0.2، متضاد 0.3 ، ارتباط عمومیت/ خاص بودن بین کلمه را 0.2 در نظر می‌گیرد واگر ارتباط R بین دو کلمه W_1 و W_2 برقرار باشد وزن ارتباط بر اساس معنای آن دو کلمه به صورت ذیل می‌باشد.

$$fs(w_1, w_2, r) = Assoc(w_1, w_2) \times DepthScore(w_1, w_2) \times RelationWeight(r)$$

امتیاز زنجیره لغوی C_i مجموع امتیاز هر یک از ارتباطات موجود r_j در زنجیره لغوی خواهد بود و بر این اساس به صورت ذیل محاسبه می‌شود:

$$Score(C\_i) = \sum\_(r\_j \in C\_i) f\_(s(w\_j1), w\_j2, r\_j)$$

روش ارائه شده بر اساس فرضیات ذیل ایجاد شده است:

فرض اول: تمامی کلمات موجود در یک زنجیره یک معنا را بیان می‌دارد.

فرض دوم: تعداد ارتباطات کلمات در یک زنجیره درجه پیوستگی را بیان می دارد. هر چقدر زنجیره اتصالات بیشتری داشته باشد، مهم‌تر خواهد بود.

فرض سوم: هر چقدر کلمه‌ای در زنجیره متصل‌تر باشد، در زنجیره مهم‌تر خواهد بود. بر این اساس فرمول جدیدی برای وزن‌دهی به عبارات به نام ConceptFreq به قرار ذیل معرفی شده است:

$$ChainFreq(w_{ij}, c_j) = wordConnectivity(w_{ij)} \times ChainConnectivity(c_j)$$

$$wordConnectivity(w_{ij}) = Frequency(w_{ij}) \times \frac{r(w_{ij}, c_j) + \alpha}{\sum_{j=1}^{l} \sum_{k=1}^{N} r(w_{kj}, c_j)}$$

$$ChainConnectivity(c_j) = \frac{\sum_{k=1}^{N} r(w_{kj}, c_j)}{|c_j|}$$

که در آن $r(w_{ij},c_j)$ تعداد ارتباطاتی است که شامل w_{ij} در زنجیره c_j می‌باشد.$|c_j|$ تعداد کلمات بی‌همتا در زنجیره c_j می‌باشد و α عدد کوچک ثابتی می‌باشد. ConceptFreq ضرب آن‌ها خواهد بود.

WordConnectivity(wij) تابعی برای اندازه گیری درجه اتصال کلمه w_{ij} در زنجیره c_j می‌باشد و ChainConnectivity(cj) درجه اتصال زنجیره c_j می‌باشد. امتیاز هر جمله‌ای مجموع وزن کلمات تشکیل دهنده آن خواهد بود که در ادامه امده است. N تعداد کلمات می‌باشد و C تعداد زنجیره‌ها است.

$$SentenceScore(Sentence\_k) = \sum\_(i = 1)^N \sum\_(j = 1)^c [\![ConceptFreq(w\_ij, c\_j)]\!]$$

ارزیابی آن بدین صورت انجام گرفته است که 20 سند به صورت تصادفی از DUC 2001 انتخاب شده‌اند و نتایج آن از روش‌های سنتی همانند tf-idf بهتر می‌باشد. ارتباطات معنایی که در WordNet موجود نمی‌باشد مشکل ساز خواهد بود.

7-2-2-2- روش‌های بر پایه گراف

-Text Rank [40] روش امتیازدهی برای پردازش متن است که بر پایه تئوری گراف می‌باشد. روش غیر نظارتی برای استخراج کلمه کلیدی و جمله ارائه کرده است. امتیازدهی بر پایه گراف بدین صورت است که امتیاز دهی به رئوس بر اساس اطلاعات سراسری و به صورت بازگشتی می‌باشد. مزیتی که این روش دارد آن است که امتیازدهی را روی تمام جملات انجام می‌دهد و بنابراین می‌توان خلاصه خیلی کوچک هم ایجاد نمود. مزیت دیگر آن این است که فقط روی متن مورد نظر پردازش انجام می‌دهد همان طور که افراد نیز هنگام خلاصه دستی همین کار را انجام می‌دهند. TextRank گرافی ایجاد می‌نماید که متن را نمایش می‌دهد و کلمه‌ها و دیگر موجودیت‌های متن را بااستفاده از ارتباط معناداری به هم مرتبط می‌نماید. برای هر جمله یک راس در نظر گرفته می‌شود. برای مرتبط کردن رئوس به یکدیگر، معیار شباهت تعریف نموده است. این ارتباط بین دو جمله را به نوعی می‌توان همان پروسه پیشنهاد در نظر گرفت. جمله‌ای با مفهوم خاصی در متن به خواننده پیشنهادی برای مراجعه به جمله دیگری در متن می‌دهد که هر دو جمله مفهوم مشترکی در متن دارند و بنابراین می‌توان لینکی بین این دو جمله در نظر گرفت. اگر Out(Vi) مجموعه رئوسی باشد که Vi به آن‌ها اشاره دارد و In(Vi) مجموعه رئوسی باشد که به Vi اشاره دارد و d پارامتری بین صفر و یک باشد. امتیاز PageRank که برای Vi در نظر گرفته می‌شود برابر است با:

$$PR(V_i) = (1-d) * \sum_{V_j \in In(V_i)} \frac{PR(V_j)}{|Out(V_j)|}$$

الگوریتم با یک مقدار اختیاری برای هر یک از رئوس انتخاب می‌شود و محاسبات تکرار می‌شود تا به یک همگرایی پایینتر از یک حد خاصی برسد. بعد از اجرای الگوریتم امتیازی به هر یک از رئوس داده می‌شود که درجه اهمیت آن راس را بیان می‌دارد. مقدار اولیه‌ای که به رئوس داده می‌شود تاثیری روی امتیاز نهایی ندارد.

- در [41] کلمه‌ها را به صورت گره‌های گراف در نظر می‌گیرد. کلمه‌هایی که بیشترین میزان ارتباط را با کلمه‌های مهم دارند، مهم‌تر هستند. منظور از ارتباط آن است که همپوشانی محتوایی و لغوی بیشتری با دیگر جملات داشته باشند. مهم‌تر بودن به وسیله الگوریتم بازگشتی pageRank [42] که گوگل نیز برای وزن دهی به صفحات وب از آن استفاده می‌کند می‌باشد.

-[43] روشی برای تخمین وزن عبارت در سند ارائه نموده است و نشان داده که این مساله برای طبقه بندی متون، کارا می‌باشد. این روش از خصوصیت هم مکانی عبارت به عنوان معیاری برای وجود وابستگی بین خصوصیات کلمه استفاده می کند. روش حرکت تصادفی در گراف کلمات و وابستگی هم مکانی کلمات اعمال می‌شود و امتیازی برای هر کلمه به وجود خواهد آمد که چگونگی مشارکت خصوصیت خاصی از کلمه در متن داده شده را بیان می‌نماید. در حرکت تصادفی اهمیت هر عبارت با احتمال آن که خواننده فرضی (حرکت کننده) به عبارت مورد نظر در متن در حین حرکت برسد، بیان می‌شود. در این معیار جدید، مکان عبارت و ارتباط آن با عبارت‌های همسایه‌اش در وزن دهی به عبارت موثر است و بر این اصل استوار است که کلمه‌های هم مکان خصوصیات مشترکی دارند. از روش امتیاز دهی به گراف به نام Page rank [42] و الگوریتم پردازش متن آن به نام TextRank [44] استفاده نموده است.

در الگوریتم حرکت تصادفی ایده اصلی بر اساس رای دادن و پیشنهاد دادن است. زمانی که یک راس به راس دیگر در ارتباط باشد، رایی را به راس دیگر می‌اندازد. هر اندازه که تعداد رای‌های بیشتری برای یک راس ارسال شود، درجه اهمیت آن بیشتر است. علاوه بر آن، راس ارسال کننده رای نیز بیانگر اهمیت آن رای خواهد بود. امتیاز هر راس بعد از هر تکرار بر اساس وزن‌های جدید رئوس همسایه دوباره محاسبه می‌شود. الگوریتم هنگامی که به نقطه همگرایی برای تمامی رئوس رسید متوقف می‌شود که به معنای آن است که نرخ خطا برای هر راسی به زیر حد خاصی رسیده است. این نحوه امتیازدهی بر اساس مدل حرکت تصادفی می‌باشد که حرکت کننده قدم‌های تصادفی بر می‌دارد. از توسعه یافته خطی مشی PageRank استفاده نموده است.

وزن یال بین دو راس V_1 و V_2 با فرمول ذیل محاسبه می‌شود:

$$Ev_1, v_2 = tf.idf_{V_1} * tf.idf_{V_2}$$

که در آن Ev_1,v_2 یالی است که V_1 و V_2 را به یکدیگر مرتبط می‌نماید. فاکتور تضعیف تابعی از وزن یال ورودی است و به قرار ذیل محاسبه می‌شود:

$d_{E_{v_1,v_2}} = E_{v_1,v_2} / \mathrm{E_{max}}$

$\mathrm{E_{max}}$ بالاترین وزن یال در گراف می‌باشد. فرمول امتیاز دهی به گره به صورت ذیل خواهد بود:

$$\dot{S}(V_a) = \frac{(1-d)}{|N|} + \sum_{V_b \epsilon In(V_a)} C * \frac{d_{E_{v_1,v_2}} * S(V_b)}{|Out(V_b)|}$$

N تعداد کل گره‌های گراف می‌باشد. d عدد ثابت تضعیف می‌باشد. C ثابت مقیاس دهی می‌باشد.

در بخش وزن دهی به عبارت مراحل ذیل برای وزن دهی انجام می‌شود:

ابتدا سند نشانه گذاری سمبول‌های خاص و کلمات خلاصه می‌شود. کلمات معمول حذف می‌شود. متن منتج پردازش می‌شود و تعداد تکرار عبارات و وزن‌های حرکت تصادفی برای هر عبارت محاسبه می‌شود. این دو خصوصیت به عنوان خصیصه‌های طبقه بندی کننده در نظر گرفته شده است. نتایج خود را با سه طبقه بندی کننده neive bayes و Rocchio و SVM مقایسه نموده است. و هم مکانی را برای پنجره‌ها با اندازه‌های 2و4 و 6 انجام داده است و نتایج بهتری حاصل شده است.

-[40] از الگوریتم TextRank استفاده نموده است و تغییراتی در امتیاز دهی به کلمات و جملات قائل سده است. اگرفرکانس کلمات را نیز در نظر بگیرد، امتیازدهی به رئوس به قرار فرمول ذیل می‌باشد:

$$PR^w(V_i) = (1-d) + d * \sum_{v_j \in In(v_i)} w_{ji} \frac{PR^w(V_j)}{\sum_{v_k \epsilon Out(v_j)} w_{kj}}$$

همپوشانی بین دو جمله Si و Sj که هر جمله با Ni کلمه نمایش داده می‌شود، Si=W1W2….WNi به صورت ذیل تعریف می‌شود:

$$Similarity(S_i, S_j) = \frac{\left|w_k | w_k \in s_i \& w_k \in s_j\right|}{\log(|s_i|) + \log(|s_j|)}$$

گرافی که بر این اساس ایجاد می‌شود یال‌های زیادی دارد و هر یال وزنی دارد که میزان قوت اتصال دو جمله را بیان می‌دارد. پس از اعمال الگوریتم امتیازدهی روی گراف، جملات به ترتیب نزولی مرتب می‌شوند و به ترتیب برای خلاصه انتخاب می‌شوند. برای ارزیابی آن از 567 مقاله که در DUC آماده شده بود استفاده شده و از ابزار ارزیابی ROUGE استفاده شده است. این روش کاملا غیر نظارتی می‌باشد و در DUC 2002 با بالاترین کارایی انتخاب شده است و این نتایج بیانگر آن است که روش بر پایه گراف PageRank که قبلا برای آنالیز لینک صفحات وب استفاده می‌شده ابزار خلاصه‌سازی مناسبی می‌باشد.

7-2-2-3- روش‌های بر پایه هم مکانی

-در[45] وزن دهی به جملات بر اساس نزدیکی به ایده اصلی متن مشخص می‌نماید که با تحلیل ساختار کلامی متن حاصل خواهد شد. [46] از وجود یک سری کلمات خاص برای وزن دهی به جملات استفاده می‌نماید. ولی این روش خلاصه را وابسته به زبان و حوزه خاص می کند.

-[47] از اطلاعات توالی متن برای تشخیص بخش‌های مناسب متن برای خلاصه استفاده کرده است. در این روش تمامی n کلمه یا همان n گرم (چند وزنی) را استخراج می‌نمایند و آن‌هایی که بیشترین توالی تکرار را در هر سندی دارند در نظر گرفته خواهند شد. چند- وزنی به معنای آن است که کلمه‌ها به یک ترتیب پشت هم در متن ظاهر شوند.تعداد دفعات تکرار عبارات چند وزنی را به عنوان معیار اهمیت آن مطرح نموده است.

- [48] روش آماری ارائه نموده است که وابسته به حوزه و زبان خاصی نمی‌باشد. با نتایج تجربی که در مقاله ذکر شده‌است است نشان داده است که کلماتی که بخشی از دو-وزنی می‌باشند و بیش از یک بار در سند تکرار شده اند، عبارت مناسبی برای بیان محتوای متن می‌باشند. همچنین در این مقاله نشان داده است که در نظر گرفتن فرکانس جمله به عنوان وزن جمله معیار مناسبی می‌باشد. چند- وزنی در صورتی تعدد تکرار دارد که از یک حدی در متن بیشتر تکرار شده باشد که در این صورت به آن‌ها چند وزنی های با تکرار بالا گویند. ایده‌ای که در این‌جا مطرح شده آن است که چنین چند- وزنی می‌تواند موضوع و ایده اصلی سند را بیان کند. چند – وزنی می‌تواند زیر مجموعه‌ای از چند-وزنی دیگری باشد. بیشینه تکرار چند-وزنی، چند-وزنی است که از یک میزان بیشتر تکرار شده و زیر مجموعه ای از چند-وزنی دیگری نیست.

روش مورد نظر خود را با روش ساده‌ای که به ترتیب از جمله‌های ابتدایی برای ایجاد خلاصه استفاده می‌شده است و همچنین با روش [44] مقایسه نموده و نتایج بهتری خاصل شده است.

- [49] این روش بر اساس خصوصیت هم مکانی کلمه پایه گذاری شده است که اگر دو کلمه در یک پنجره مکانی مشاهده شوند، از نظر مفهومی به یکدیگر مرتبط خواهند بود. هر چقدر که این درجه هم مکانی بیشتر باشد، به یکدیگر بیشتر مرتبط خواهند بود. روش بیان شده در [50] اصلاح شده است و با خصوصیات مفهومی نیز ترکیب شده است. تفاوت‌هایی که این روش دارد به قرار ذیل می‌باشد:

1) روشی که برای انتخاب کلمات مهم و کلمات ارتباط دهنده استفاده شده است.

2) روشی که برای انتخاب جمله از هر کلاستر به کار رفته است.

3) به کارگیری ارتباط بین کلمات

همانند [50] درجه هم مکانی کلمات به صورت ذیل تعریف می‌شود:

$$R(w_i|w_j) = \frac{f(w_i,w_j)}{f(w_i)}$$

$f(w_i)$ فرکانس w_j و $f(w_i, w_j)$ تعداد دفعاتی می‌باشد که این کلمات در یک پنجره مشاهده شده‌اند. در نظرگرفتن بهترین اندازه پنجره برای محاسبه درجه هم مکانی بسیار مهم می‌باشد و روی دقت متن خلاصه تاثیر خواهد گذاشت. در این بررسی بهترین پنجره مکانی، متوسط طول جمله در نظر گرفته شده است. درجه هم‌مکانی همانند [50] به صورت ذیل تعریف می‌شود:

$$C(w_i, w_j) = \frac{R(w_i|w_j) + R(w_j|w_i)}{2}$$

برای هر دو کلمه، درجه هم‌مکانی محاسبه می‌شود. زنجیره لغوی بر اساس پایگاه داده لغوی کلمات هم‌معنی ایجاد می‌شود. از آنجایی که یک کلمه می‌تواند معانی مختلفی داشته باشد و علاوه بر آن به علت نبود صدا در زبان فارسی یک کلمه می‌تواند متعلق به یک یا بیشتر زنجیره لغوی باشد. برای آن‌که زنجیره لغوی مناسب انتخاب شود، به این صورت عمل می‌شود که مجموع درجات هم مکانی هر یک از زنجیره‌های لغوی محاسبه و نهایتا زنجیره لغوی با بیشترین مجموع، انتخاب خواهد شد. اگر زنجیره لغوی مرتبطی وجود نداشته باشد، یک زنجیره تغوی جدید ایجاد خواهد شد. معادله3 این مساله را بیان می‌دارد

$$Chain(w_{new}) = Chain_w$$
$$where: S_{Chain_w}(w_{new}) is \boldsymbol{Maximum} \text{ and } S_{Chain_w}(w_{new}) = \sum_{w \in Chain_w} C(w_{new}, w)$$

امتیاز هر کلمه همانند آن‌چه در معادله 4 آمده است از طریق ضرب فرکانس هر کلمه در تعداد کل زنجیره‌های مرتبط به دست می‌آید:

$$Rank(w) = Frequency(w) \times WordCount(Chain_w)$$

N کلمه با امتیاز بالاتر انتخاب خواهند شد که در این‌جا n ، برابر تعداد کل کلمات سند تقسیم بر 20 انتخاب شده است. بر این اساس گرافی تشکیل می‌شود که کلمات مهم در آن راس‌ها می‌باشد و در صورتی یال بین دو راس در نظر گرفته خواهد شد که درجه هم‌مکانی بین آن‌ها از یک حد پیش تعریف

شده‌ای بیشتر باشد. بر اساس تجربه 0.7 بهترین حد ممکن به نظر می‌آید. در صورتی که متن ورودی به بیش از یک موضوع مرتبط باشد، گراف نامتصل خواهد بود.

حال برای آن‌که پیوستگی متن محفوظ شود، کلماتی که متعلق به بیش از یک کلاستر باشند، مشخص می‌شوند. این کلمات از طریق محاسبه احتمال تعلق یک کلمه به بیش از یک کلاستر انتخاب خواهند شد. احتمال آن‌که یک کلمه به بیش از یک کلاستر C_i تعلق داشته باشد همانند[50] تعریف شده‌است:

$$p(w \in C_i) = \frac{\sum_{\substack{w_c \in C_i \\ w \neq w_c}} C(w,w_c)}{\sum_{k=1}^{ClusterCount} \sum_{\substack{w_c \in C_k \\ w \neq w_c}} C(w,w_c)}$$

که مجموع درجه هم‌مکانی یک کلمه به هر یک از کلمه های کلاستر تقسیم بر مجموع وابستگی کلمه به تمامی کلاسترها می‌باشد. احتمال آن‌که یک کلمه به بیش از یک کلاستر متعلق باشدبه صورت ذیل خواهد بود:

$$LinkScore(w) = 1 - \sum_{i=1, j \neq i}^{ClusterCount} (p(w \in C_i) \times p(w \notin C_j))$$

که در آن Count تعداد کل کلاسترها می‌باشد. حال m کلمه با بالاترین linkscore انتخاب و به گراف افزوده می‌شود. برای هر کلمه امتیازی از طریق جمع وزن هر یک از یال‌های متصل حساب خواهد شد.

امتیاز هر جمله نیز از طریق جمع امتیاز کلمات آن محاسبه خواهد شد. در این محاسبه، طول جمله نیز در نظر گرفته می‌شود. امتیاز جمله از طریق ذیل محاسبه می‌شود:

$$w(S_i) = \frac{\sum w_i}{|S_i|}$$

از آن‌جایی که تجربه نشان داده است که جمله اول در بیشتر متون مهم می‌باشد، در خلاصه خواهد آمد.

بعد از امتیاز دهی به جملات، n جمله با بالاترین امتیاز انتخاب خواهند شد. ولی می‌بایست توجه شود که تعداد جملاتی که از هر یک از موضوعات متن در خلاصه می‌اید متناسب با تعداد کل جملات هر موضوع باشد. برای این‌کار ضریبی متناسب با تعداد جملات تشکیل دهنده هر کلاستر در نظر گرفته می‌شود. از این ضریب بعدا برای انتخاب جملات هر کلاستر استفاده خواهد شد. ضریب خلاصه‌سازی به صورت ذیل خواهد بود:

$$Portion(C_i) = \frac{WordCount(C_i)}{TotalNumberOfWords}$$

درصد خلاصه‌سازی متغیر می‌باشد و مطابق با درخواست کاربر قابل تغییر می‌باشد. نتایج نشان می‌دهد که اگر درصد خلاصه‌سازی 30% باشد، در حدود 70%-80% درصد در یک متن خبری 3-4 صفحه‌ای وجود خواهد داشت [36].

در مرحله آخر، شباهت بین جملات از طریق معیار cosine محاسبه خواهد شد. برای هر جمله برداری از کلمه ایجاد خواهد شد . شباهت بین جملات در صورتی که دارای کلمات یکسان و یا دارای کلمات هم معنی باشند (در یک زنجیره لغوی قرار داشته باشند) افزوده خواهد شد. پس از آن جملاتی که شباهتی بیش از یک حد از پیش تعریف شده‌ای داشته باشند، حذف خواهند شد و جمله با طول بیشتردر خلاصه باقی‌خواهد ماند. با این‌کار افزونگی در متن خلاصه وجود نخواهد داشت.

از آن‌جایی که benchmark شناخته شده‌ای در زبان فارسی برای ارزیابی روش پیشنهادی وجود ندارد، نتایج با FarsiSum مقایسه شده است که تنها خلاصه‌ساز فارسی می‌باشد. از 60 متن خبری و داستانی به عنوان مجموعه تست استفاده شده است. برای هر متن، 5 دانشجو کار خلاصه‌سازی را به طور دستی انجام داده‌اند. این کار در 3 نسبت متفاوت نسبت به متن اصلی30%, 40% و 50% انجام شده‌است. خلاصه ایده‌آل بر اساس رای‌گیری اکثریت از نتایج دستی انجام می‌شود. Precision و recall مربوط به روش پیشنهادی و FarsiSum محاسبه شده است. این نتایج در جدول 1 آمده‌است و نشان می‌دهد که تکنیک مطرح شده خصوصا زمانی که نرخ فشردگی 30% باشد، بهتر عمل می‌نماید.

جدول 1- نتایج ارزیابی

نرخ فشردگی	پارامتر	روش پیشنهادی	FarsiSum
30%	Precision	0.64	0.57
	Recall	0.61	0.55
40%	Precision	0.62	0.54
	Recall	0.61	0.52
50%	Precision	0.62	0.53
	Recall	0.60	0.52

- در [51] از چند وزنی به عنوان یکی از خصوصیات جمله در روش یادگیری غیر نظارتی استفاده نموده است.

- در[52] [18] نیز از توالی کلمات برای انتخاب جملات و تولید خلاصه استفاده نموده‌اند.

- در [53] کلماتی که از نظر نحوی به یکدیگر اتصال دارند به عنوان عنصر پایه‌ای برای تحلیل متن درنظر گرفته می‌شود. نیاز به پیمایش درخت وابستگی نحوی دارد. این روش از 4 بخش اصلی تشکیل شده است:

1) پیش پردازش: در این مرحله حذف کلمه‌های توقف و پس از آن ریشه یابی انجام می‌شود
2) انتخاب عبارات: تمامی تعداد دفعاتی که n کلمه پشت سر هم در متن ظاهر شده‌اند را در نظر می‌گیرد. به عبارت دیگر چند- وزنی‌ها را استخراج می‌نمایند و از روی بیشترین توالی تکرار چند وزنی که زیر مجموعه چند وزنی دیگری نباشد مقدار n مشخص خواهد شد. تعداد تکرار هر عبارت چند وزنی از طریق تعداد دفعاتی که n کلمه به دنبال یکدیگر در سند ظاهر شده اند به دست می‌آید. مزیت این روش آن است که n را ثابت درنظر نگرفته است و بسته به هر سند، n ممکن است تغییر نماید. به عبارت دیگر یک توالی کلمه بیشینه که به تعداد زیاد در متن تکرار شده است انتخاب خواهد شد.
3) وزن دهی به عبارات: بر اساس هر 4 روش ذیل صورت گرفته‌است و در مورد هر یک ارزیابی جداگانه انجام شده است:

- وزن دهی منطقی: اگر کلمه در سند باشد مقدار یک و در غیر این صورت مقدار آن صفر خواهد بود
- فرکانس کلمه: بر اساس این اصل ساده است که کلمه‌ای که در یک سند موجود است بهتر می‌تواند محتوای سند را بیان کند [54]. بنابراین TF امتیاز بیشتری به کلماتی که بیشتر تکرار می‌شود، می‌دهد. $W_i(t_j)=f_{ij}$
- فرکانس معکوس سند: سالتون این روش را ارائه کرد [55]. بدین مضمون که وقتی کلمه‌ای در تمامی سندها ظاهر می‌شود این کلمه برای تفکیک و تشخیص اسناد از یکدیگر بلااستفاده می‌باشد. $w_i(t_j)=\log(N/n_j)$ که N تعداد کل اسناد می‌باشد و n_j تعداد اسنادی می‌باشد که کلمه j در آن‌ها وجود دارد.

- فرکانس کلمه/ معکوس فرکانس سند: مشکل فرکانس معکوس سند آن است که امکان متمایز کردن دو سند که کلمات یکسان دارند وجود ندارد حتی اگر کلمه‌ای در یکی از اسناد بیشتر تکرار شده باشد. این معیار کلماتی را که در یک سند بیشتر تکرار شده اند ولی در کل اسناد کمتر تکرار شده‌اند را مرتبط تر محسوب خواهد کرد که فرمول آن به قرار ذیل می‌باشد:

$w_i(t_j)= f_{ij} * \log(N/n_j)$

4) انتخاب جملات بر اساس یادگیری غیر نظارتی

بعد از امتیاز دهی به جملات از آن‌جا که ممکن است به عنوان مثال جمله دوم به جمله اول شباهت زیادی داشته باشد، گروه جملات مشابه هم در نظر گرفته می‌شود تا بدین صورت جمله افزونه نداشته باشیم. کلمه‌های هم معنی یکسان در نظر گرفته شده‌است.

علاوه بر آن از TF-IDf برای وزن دهی به عبارات انتخاب شده استفاده می‌شود. در مرحله انتخاب جملات، از الگوریتم k-mean برای کلاستر کردن جملات مشابه استفاده می‌شود تا بدین صورت از هر گروه یک جمله مشخص شود. بر اساس نرخ خلاصه‌سازی تعداد کلاسترها را خودش انتخاب می‌نماید. زیرا همان‌طور که می‌دانیم در k-mean تعداد کلاسترها می‌بایست مشخص باشد. به عنوان مثال اگر متوسط طول جمله را 20 کلمه بگیریم وکاربر خلاصه 100 کلمه‌ای بخواهد، 5 کلاستر می‌بایست ایجاد شود. هر سند بر اساس برداری از چند-وزنی ها نمایش داده می‌شود. k-mean یک سری نقاط میانی می‌گیرد و هر کدام از جملات را به نزدیکترین کلاستر منتصب می‌نماید.و مجددا در هر کلاستری نقاط میانی محاسبه می‌شود تا دیگر نقطه میانی عوض نشود. در این صورت الگوریتم متوقف می‌شود. K-mean در صورتی جواب درست می‌دهد که نقاط ابتدایی به خوبی انتخاب شوند. جمله اول را می‌توان به عنوان seed انتخاب نمود. داده تست DUC بوده است. بعد از آن‌که الگوریتم k-means اتمام یافت، نزدیکترین جمله به هر مرکزی را برای تولید خلاصه انتخاب می‌نماید.

برای ارزیابی 570 مقاله خبری با اندازه‌های مختلف و از موضوعات مختلف انتخاب نموده است. هر کدام از این اسناد در DUC توسط دو فرد خبره خلاصه شده است. در مورد ارزیابی از ابزار ارزیابی ROUDGE استفاده نموده است که دو خلاصه اتوماتیک و دستی را با استفاده از چند- وزنی با هم مقایسه می‌نماید. نتایج ارزیابی نشان داده است که سه- وزنی مناسب می‌باشد.

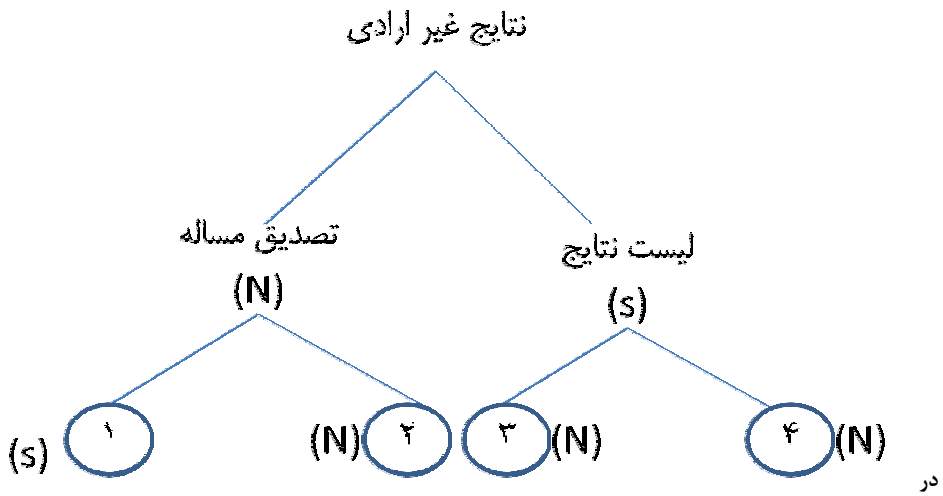
7-2-3- روش‌های بر پایه ساختار کلامی

- در روشی که [56] ارائه نموده است، ریشه درخت تئوری ساختار زبانی امتیازی معادل تعداد سطوحش دارد. پس از آن درخت به روش عمقی پیمایش می‌شود و هر زمانی که به بخش غیر مهم می‌رسد امتیاز را یکی کم می‌نماید. بخش‌های متن بر اساس امتیازشان در درخت مرتب می‌شوند.

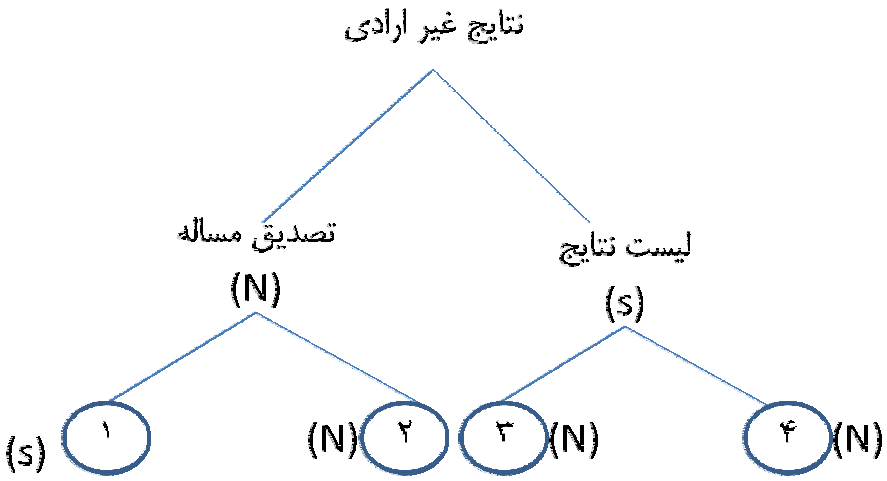
- در [6] مجموعه پیشرفت برای هر گره درخت به این صورت تعریف شده است که برابر اجتماع فرزندان پایه اش می‌باشد. برای امتیاز دهی به هر بخش، پایه بودن آن، اهمیت ارتباط و مجموعه پیشرفت آن در نظر گرفته شده است. به ریشه درخت امتیازی معادل دو برابر تعداد سطوح درخت داده شده است. برای امتیاز دهی به هر بخش، درخت پیمایش می‌شود. هر گره یا بخشی که به مجموعه پیشرفت تعلق نداشته باشد، امتیازش با متمم ضریب اهمیت ارتباط کم می شود. نهایتا بین هر بخشی با جمله اصلی متن شباهت cosine محاسبه می‌شود و در امتیاز نهایی هر بخش ضرب می شود. به عنوان مثال در 2 جمله ذیل:

1) اگر چه که نسبت به آن آلرژی داشت، (2) آن را امتحان کرد.

(3) اکنون سر درد داردو (4)بدنش قرمز شده است.



در شکل 1 درخت ساختار زبانی بالا آمده است.



شکل 1- مثالی از درخت ساختار کلامی

N به معنی پایه جمله و حاوی اطلاعات مهم است. S پیرو و جز فرعی اطلاعات را نشان می‌دهد. این مساله که شخصی آلرژی دارد باید باعث شود که آن چیزی که آلرژی دارد را امتحان ننماید که این، بخش اصلی متن است.نتایج چنین کاری جمله 3 و 4 است که هر دو نهاد هستند و بخش فرعی ندارد. ارزیابی آن بدین صورت انجام پذیرفته است که 30 متن که خلاصه دستی اشان نیز موجود می‌باشد را انتخاب کرده‌اند و precision و recall محاسبه شده است.

- [45] روشی برای تجزیه کلامی و خلاصه‌سازی ارائه نموده است که انسجام و پیوستگی متن را حفظ می‌نماید. ساختار کلامی ایجاد نموده است که شبیه درخت تئوری ساختار بیانی [15] می‌باشد با این تفاوت که این درخت دودویی می‌باشد و نام ارتباطات در آن موجود نمی‌باشد. ساختن درخت کلامی شبیه روش بر پایه کلمات خاص است [23] ولی محدودیت‌هایی که به خاطر وجود ارتباط بین ساختار کلامی و زنجیره‌های مورد رجوع (آشکار سازی انسجام) از طرفی و ساختار کلامی سراسری (آشکارسازی پیوستگی) از طرف دیگر در تئوری vein آمده است را نیز در بر می‌گیرد[57] .ماجول پیشنهادی شامل بخش‌های برچسب‌سازی اجزای کلام (part of speech)، تجزیه کننده FDG، بخش بندی کردن اجزای جمله، تشخیص اسامی، تفکیک کردن بخش‌هایی که تکرار می‌شوند، تجزیه کلامی و خلاصه‌سازی است.

تمرکز این روش روی خلاصه‌هایی که در بازیابی اطلاعات اخبار یا مقاله‌های علمی که کاربر موضوع خاصی را جستجو می‌نماید، می‌باشد

ابتدا متن با استفاده از سیستم نقش کلمه برچسب گذاری می‌شود. پس از آن تجزیه کننده نحوی FDG روی آن اجرا می‌شود. پس از آن دو پردازش دیگر انجام می‌شود. اولین پردازش جمله‌ها را به واحدهای کلامی ابتدایی تبدیل می‌نماید (edus) و سپس درخت کلامی ابتدایی (edts) را ایجاد می‌نماید و دیگری عبارات اسمی را تشخیص داده و از موتور تفکیک کردن عبارات تکراری استفاده می‌نماید. بنابراین هر edt درخت کلامی می‌باشد که برگ‌های آن edus یک جمله می‌باشند. برای هر جمله در متن اصلی مجموعه‌ای از edts ها به دست می‌آید. در این مرحله پردازش کلامی افزایشی شروع می‌شود. بعد از n مرحله متناظر با n جمله اول، جنگلی از درخت‌ها وجود خواهد داشت که نشان دهنده ترکیب edt های n جمله می‌باشد. بنابراین هر درختی یک تفسیر از متن را بیان می‌دارد. در مرحله n+1 در پردازش کلامی افزایشی عملیات ذیل انجام خواهد شد:

ابتدا تمامی edt های متناظر با جمله بعدی با تمام حالات ممکنه به جنگل موجود مجتمع می‌شود. سپس درخت‌های نتیجه امتیازدهی و مرتب و فیلتر می‌شوند و بنابراین تنها بخشی از آن‌ها باقی می‌ماند. از درخت‌های باقی مانده که از مرحله نهایی بدست می‌آید، درخت‌های با بالاترین امتیاز انتخاب می‌شوند و خلاصه بر اساس درخت‌ها ایجاد می‌شود.

7-2-4- روش‌های ترکیبی

در یک جمع بندی از آن‌جایی که تحلیل متن به صورت عمیق روی متن‌هایی که محدود به یک حوزه خاص نباشد مشکل است، بنابراین روش‌هایی که بر پایه هوش مصنوعی می‌باشند، در حوزه‌های محدودی می‌توانند عمل نمایند. تمایل به این سمت وجود دارد که از هر دو این روش‌ها به نحوی استفاده شود.

7-2-4-1- روش‌های بر پایه آموزش یادگیری

خلاصه‌سازی متن به عنوان یک مساله طبقه‌بندی نیز می‌تواند در نظر گرفته شود [46, 58]. در فاز یادگیری، داده آموزش که در اینجا همان جمله می‌باشد به الگوریتم طبقه بندی داده می‌شود و داده‌های تست (جمله‌های دیده نشده) به دو گروه طبقه‌بندی می‌شوند، آن‌هایی که در خلاصه شامل می‌شوند و آن‌هایی که در خلاصه مشاهده نمی‌شوند. برای طبقه‌بندی می‌بایست خصوصیاتی از متن مورد توجه

قرارگیرد. [59] از الگوریتم ژنتیک برای استخراج خصوصیاتی از متن استفاده می‌کند ولی دقت روش در این حالت 52% می‌باشد.

- خلاصه‌ساز فارسی در [60] ارائه شده است که از یک روش ترکیبی آماری و معنایی استفاده نموده و شباهت جملات با یکدیگر، کلمات کلیدی و با عنوان را مشخص می‌نماید. ابتدا اسامی مفید تشخیص داده می‌شوند. مرز جمله‌ها با استفاده از علائم جدا کننده‌ی جمله شامل (".","؛"،"!"،" ؟"، "?" و :") و مرز کلمه‌ها با استفاده از علایم فضای خالی،" ، ","،"،">"،"<"،"]"،"["،"-"،""،"/"و خط جدید مشخص می‌شوند. منبعی شامل کلمه‌های غیر مهم را برای حذف کلمه هایی که زیاد استفاده می شوند اما مهم نیستند ایجاد شده‌اند. کلمات مهم، اسامی و صفات هستند که ریشه‌ی هر کدام را با الگوریتم ریشه یابی می‌کنند و پایگاه داده از اسامی مرکب هم وجود دارد. سپس گرافی تشکیل می‌دهد که در آن گراف، هر جمله یک گره است و یال‌ها ارتباط بین جملات را بر اساس زنجیره‌های لغوی بیان می‌کنند جملاتی که میزان ارتباط آن ها از آستانه‌ی خاصی بیشتر است با یالی به یکدیگر متصل می شوند و میزان این ارتباط وزن یال را مشخص می کند. الگوریتم ایجاد زنجیره‌ها روی ریشه‌ی کلمات به روش زیر اعمال می‌شود:

هر ریشه‌ی s در واژگان مفهومی جستجو می‌شود و ریشه‌ی تمام کلماتی را که s با آنها رابطه دارد و نوع رابطه‌ی آنها یافته می‌شود. به فرض مجموعه‌ی $W= \{w_1,w_2,\ldots,w_n\}$ شامل ریشه‌ی این کلمات باشد. تمام w_i ها را در زنجیره های از قبل موجود جستجو و اگر در زنجیره های قبلی وجود داشتند ، شماره ی زنجیره ی آن ها یعنی مجموعه ی $C=\{c_1,c_2,\ldots,c_n\}$ یافته می‌شود که در آن c_i زنجیره‌ی مربوط به w_i است، کلماتی را که در زنجیره های قبل موجود نبودند نیز مشخص شده (مسلماً c_i ها می توانند مشترک باشند) با توجه به حریصانه بودن الگوریتمی c_i را که بیشتر از همه در مجموعه‌ی بالا تکرار شده یافته و تمام کلمات این زنجیره را بررسی می‌نماید. اگر تمام آن ها در W بودند، ریشه‌ی جدید را در همین زنجیره درج و در غیر این صورت این مرحله بدون این c_i تکرار می‌شود تا یا زنجیره ی مناسب یافته و یا هیچ زنجیره ای برای آن یافت نشود که در این صورت زنجیره‌ی جدیدی برای آن ایجاد می‌شود. بعد از ایجاد زنجیره ها میزان شباهت بین هر دو جمله از طریق ضرب برداری کلمات مفید دو جمله با در نظر گرفتن رابطه معنایی بین کلمات مشخص خواهد شد.

رابطه‌ی تکرار ، هم معنی، مخالف، تعمیم، تخصیص، اشتقاق و هم- مکانی در نظر گرفته شده و به هر کدام وزن خاصی داده و رابطه‌ی زیر بین آنها متصور شده است:

تکرار < هم‌معنی <اشتقاق = هم مکانی < مخالف< تعمیم= تخصیص

امتیاز شباهت برای هر جمله با جمع مقادیر شباهت آن با سایر جملات محاسبه می‌شود . امتیاز شباهت هر جمله با کلمات کلیدی کاربر و عنوان نیز محاسبه می‌شود. علاوه بر سه امتیاز نامبرده شباهت هر دو جمله با یکدیگر ،درجه ی هر گره در گراف که معادل با تعداد جملات مرتبط با هر جمله است نیز محاسبه می‌شود. امتیازی نیز برای جملات حاوی کلمه‌های خاص همانند بنابراین، نتیجه، موضوع و معرفی در نظر گرفته می‌شود و به جمله‌های حاوی کلماتی همانند مثال، نمونه و مثلا امتیاز منفی داده می‌شود. نهایتا امتیاز هر جمله مجموع امتیاز موارد فوق با در نظر گرفتن ضرایبی می‌باشد. برای ارزیابی به این صورت عمل شده که 10 متن از روزنامه همشهری با نسبت فشردگی 30% در نظر گرفته شده است و چند نفر دانشجوی کارشناسی ارشد به صورت دستی کار خلاصه سازی را انجام داده‌اند و جملاتی که در خلاصه‌های افراد بیشتری بوده‌است به عنوان خلاصه مرجع در نظر گرفته شده‌است. شباهت خلاصه مرجع با خلاصه سیستم 56.66 بوده‌است.

7-2-4-2- ترکیب روش سطحی و معنایی

[24] SUMMARIST در دانشگاه کالیفرنیای جنوبی مثالی از ترکیب این دو روش می‌باشد که از WORDNET [4] و خصوصیات آماری متن استفاده نموده است. ابتدا موضوع اصلی سند را بر اساس روش‌های آماری همانند تعداد کلمه‌ها و مکان آن‌ها استخراج می شود. سپس با استفاده از WORDNET و ارتباط بین کلمات جملاتی که مفاهیم عمومی‌تری از متن را دارا می‌باشند، انتخاب می‌کند.

- [61] از یک روش ترکیبی که هم مکانی کلمات را در نظر می‌گیرد استفاده نموده‌است. روش ارائه شده شامل مراحل ذیل می‌باشد:

1) پیش پردازش: در این بخش کار ریشه یابی و استخراج کلمه‌ها و فعل‌ها برای کشف ارتباط بین کلمه‌ها انجام می‌شود.
2) مشخص کردن وزن جملات: از اصل هم‌مکانی کلمات استفاده نموده است.احتمال آن‌که دو کلمه در سند هم مکان مشاهده شود. احتمال آن‌که کلمه t با k مشاهده در یک سند برابر است با:

$$P_r(k) = (1-\propto)\delta_{k0} + \left(\frac{\alpha}{B_1}\right)\left(1-\frac{1}{B_1}\right)^{k-1}(1-\delta_{k0})$$

α احتمال آن است که حداقل یک بار کلمه در سند مشاهده شود. اگر k صفر باشد، دلتای k,0 برابر یک و در غیر این صورت صفر خواهد بود. B_1 تعداد مشاهدات مورد انتظار در اسنادی است که حداقل یک‌بار مشاهده شده است.

3) کشف ارتباط بین کلمات: ارتباط اسم- اسم و اسم - فعل را در نظر می‌گیرد. بر اساس این فرض که ارتباط اسم- فعل در رابطه نهاد و گزاره‌ای و ارتباط اسم- اسم در سطح کلامی موجود است و بر اساس احتمال مشاهده اشان با یکدیگر در یک سند می‌باشد.

ارتباط اسم- اسم با فرمول ذیل تعریف می‌شود:

$$SNN(N_i) = \sum_j \frac{P_{N_i} \times P_{N_j}}{D(N_i,N_j)}$$

ارتباط اسم - فعل به صورت ذیل تعریف می‌شود:

$$SNV(N_i) = \sum_j \frac{P_{N_i} \times P_{V_j}}{D(N_i,V_j)}$$

میزان قدرت ارتباطی کلمه‌ها از جمع ارتباط اسم - اسم و اسم - فعل به دست می‌آید. نهایتا میزان مرتبط و مهم بودن جمله با میانگین گرفتن از قدرت ارتباطی کلمه‌های موجود در جمله قابل محاسبه می باشد.

4) تولید خلاصه: k جمله با بالاترین امتیاز برای خلاصه انتخاب می‌شوند. K پارامتری است که بستگی به درصد خلاصه‌سازی دارد. به عنوان مثال اگر 20 جمله در سند اصلی موجود باشد و درصد خلاصه‌سازی 50% باشد، 10 جمله انتخاب خواهند شد.

ارزیابی به این صورت انجام گرفته است که روی 800 سند از reuter کار خلاصه‌سازی و طبقه بندی را انجام داده است.

7-2-4-3- روش افزایش و کاهش جمله

- [62] برای فرایند تولید خلاصه بعد از انتخاب جملات مهم ، روشی را ارائه نموده است که در آن جملات استخراج شده ویرایش شده و با یکدیگر ترکیب می‌شوند. علاوه بر آن بخش‌هایی از جمله حذف می‌شوند. بر اساس تحلیل آماری روی مجموعه متون‌ها، دانش نحوی و محتوای جمله مشخص می‌شود چه بخش‌هایی از هر جمله حذف شود. اساس این روش بر پایه آن است که ادغام و حذف بخش‌هایی از

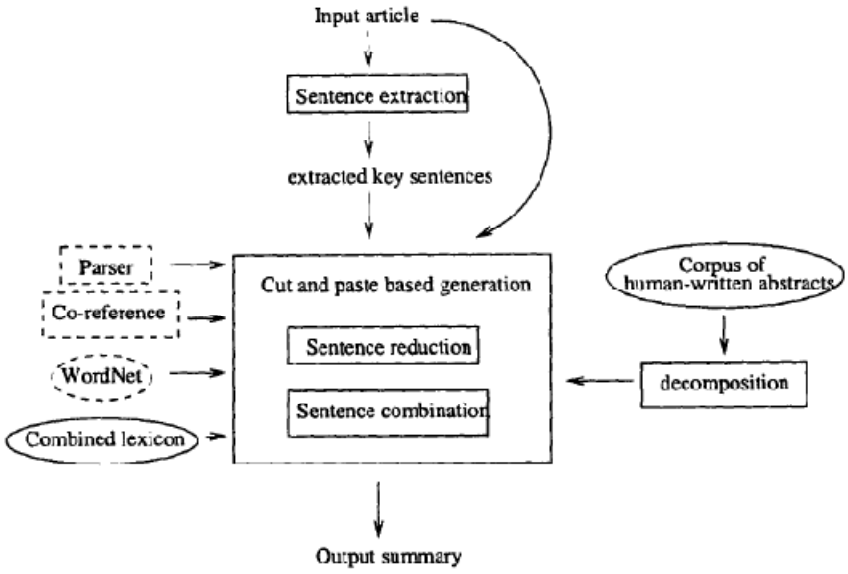
جملات، انسجام متن را بیشتر خواهد نمود. مثالی از کاهش دادن جملات، توضیحات تکمیلی، عبارات با علامت نقل قول مستقیم، معرف‌های چندگانه اسامی، مثال‌ها، حرف ربط سر جملات و تفصیل می‌باشد.

دو عمل تعریف نموده است که اساس مقایسه خلاصه انسان و ماشین می‌باشد. این دو عمل، کاهش جمله و ترکیب جملات می‌باشد. خلاصه دستی انسان را در نظر گرفته و بررسی می‌کند هر عبارتی در متن خلاصه دستی چه معادلی در متن خلاصه انسان دارد. از روش‌های یادگیری ماشینی استفاده می‌شود و مجموعه تست و آزمون دارد. نهایتا نتایجی که از آموزش و تست مجموعه بدست می‌آید مشخص می‌نماید که آیا جمله موجود در خلاصه دستی بر اساس اعمال حذف و چسباندن جملات بدست آمده است و در این صورت، چه عباراتی از جمله اصلی آمده است و در کدام قسمت سند این عبارت وجود داشته است. برای آنکه مشخص شود هر کلمه در خلاصه در چه مکانی از مقاله اصلی آورده شده است از مدل مارکوف استفاده نموده است. در حقیقت بر اساس مشاهدات، یک سری قواعد تجربی ایجاد نموده اند. از این قواعد برای ایجاد مدل مارکوف پنهان[63] استفاده شده است. کاهش جمله باید به صورتی باشد که معنای اصلی جمله را از بین نبرد. از تجزیه کننده ESG [64] استفاده نموده است تا مشخص شود که نبود عبارت کاهش یافته متن را از نظر گرامری دچار مشکل نمی‌نماید. بنابراین هر عبارتی یک امتیاز خواهد گرفت. سه نوع احتمال عبارت حذف شده، تغییر نکرده و یا تا حدودی تغییر کرده است بر اساس متن و خلاصه دستی آن تعریف می‌شود. یک عبارت در صورتی حذف می‌شود که از نظر گرامری بلامانع باشد، معنای جمله بر روی آن سوار نشده باشد و احتمال بالایی داشته باشد که توسط انسان حذف شود. دو عمل اصلی حذف و چسباندن جملات به زیر اعمال زیر تفکیک می‌شود.

1) کاهش جمله: به معنی حذف عبارات از جمله می‌باشد. عنصر حذف شده می‌تواند کلمه، عبارت یا جزئی از جمله باشد.
2) عام یا خاص کردن جمله: به معنای جایگزین کردن عبارات یا بخشی از جمله با یک توصیف عمومی‌تر یا خاص‌تر می‌باشد.
3) ترکیب جمله: عناصر چندین جمله را در هم ادغام می‌نماییم.
4) تبدیل جمله: در هر دو مورد کاهش و ادغام جمله به کار می‌آید. به عنوان مثال، مکان یک عبارت از انتها به ابتدا بیاید.
5) تفسیر لغوی: به جای استفاده از عبارات، از تفسیر آن‌ها استفاده شود.

6) مرتب کردن دوباره جملات: ترتیب جملات استخراجی را تغییر می‌دهد. مثلا جمله آخر را در اول چکیده بگذارد. در چکیده‌های دست نویس انسان، جملات عموما دوباره از اول نوشته می‌شود.

شکل 2 معماری پیشنهادی را نشان می‌دهد. ورودی یک سند می‌باشد که می‌تواند در هر حوزه‌ای باشد. در مرحله اول یا استخراج، جمله‌های کلیدی مشخص می‌شوند. مرحله دوم، ماجول کاهش و ترکیب جمله اجرا می‌شود. ورودی‌های دیگر آن چکیده‌های دستی، منابع و ابزار های دیگر، مجموعه اسناد و چکیده‌های دستی آن‌ها، تجزیه کننده نحوی، پایگاه داده لغوی و علاوه بر آن لغت نامه بزرگی که از منابع مختلف تشکیل شده است، می‌باشد. برنامه تجزیه، برای تحلیل ساختار جمله در خلاصه دستی به کار می‌رود. ماجول استخراج بدین صورت عمل می‌نماید که کلمه‌های کلیدی استخراج می‌شوند. بر اساس مجموعه ارتباطاتی که در WORDNET تعریف شده است، هر کلمه‌ای در جمله امتیازی برابر تعداد لینک‌هایی که با دیگر کلمات خواهد داشت دارد. امتیاز جمله برابر جمع امتیاز کلماتش می‌باشد. نتیجه ماژول کاهش جمله یا به طور مستقیم خلاصه را ایجاد می‌نماید یا آن‌که به ماژول ترکیب ارسال می‌شود. در بخش ترکیب جملات، ابتدا تحلیلی به صورت دستی یا با استفاده از یادگیری ماشینی روی مجموعه‌ای از جملاتی که توسط انسان ترکیب شده است انجام می‌شود و بر اساس آن مجموعه عملیات ترکیب مشخص می‌شود



شکل 2- معماری سیستم پیشنهادی

میزان precision و recall بالای 85% بوده است در قسمت برنامه تجزیه. ارزیابی نیز بدین صورت انجام داده اند که از افراد خواسته‌اند که متن دستی و متن تولید شده توسط سیستم را قضاوت نمایند.

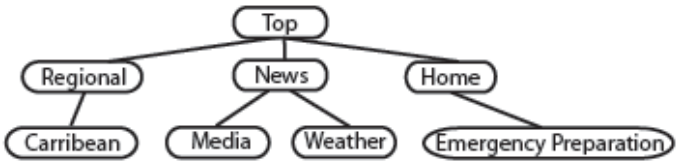
7-2-4-4- روش بر پایه هستان‌شناسی

این روش‌ها بر این اصل استوار می‌باشند که با استفاده از هستان‌شناسی معنای متن را بهتر می‌توان نشان داد.

- [65] از هستان‌شناسی یا لغت نامه مفاهیم در مورد پژشکی به نام UMLS [66] استفاده نموده است. به صورت دستی یک هستان‌شناسی برای یک حوزه خاص ایجاد نموده است و هر جمله‌ای که یک برچسب از هستان‌شناسی را در درون خود داشته باشد، امتیازش افزایش می‌یابد. هر مفهومی یک سری کلمات را شامل می‌باشد و بررسی می‌شود که هر پاراگرافی چه میزان از این کلمات را در درون خود دارد. از ارتباط عمومیت بخشی (generalization) نیز استفاده کرده تا موضوع اصلی هر پاراگرافی را بهتر بیابد. هر پاراگرافی که بالاترین امتیاز را دارد به عنوان موضوع اصلی شناخته می‌شود و انتخاب پاراگراف‌ها به ترتیب امتیازشان ادامه می‌یابد تا به میزان خلاصه‌سازی مطلوب برسد.

- در [67] هستان‌شناسی به صورت دستی برای حوزه کوچکی از اخبار ایجاد نموده است. و هر پاراگرافی که یک برچسب از هستان‌شناسی را در درون خود داشته باشد، امتیازش افزایش می‌یابد. این روش ارتباط بین کلمه‌ها حتی هم‌معنی بودن را در نظر نمی‌گیرد و واحد تولید خلاصه آن پاراگراف می‌باشد.

- در [68] این روش، جملات را به گره‌های درخت هستان‌شناسی از پیش تعریف شده طبقه بندی می‌کند. هر گره‌ای از درخت شامل مجموعه‌ای از کلمات می‌باشد. جملات به صورت زیر درخت‌هایی از هستان‌شناسی نمایش داده می‌شوند که بر آن اساس می‌توان معیار شباهت را در هستان‌شناسی استفاده نمود و ارتباط بین جملات را بر اساس خصوصیات گراف محاسبه کرد. از درخت طبقه بندی کننده DMOZ استفاده نموده است [69] و دو سطح اول آن در نظر گرفته شده که شامل 1036 گره می‌باشد. درشکل 3 مثالی از ساختار درخت DMOZ آمده است.



DocId: AP880917-0095
Tags: {Weather=0.966, Caribbean=0.585, Regional=0.563, News=0.506, Emergency Preparation=0.393, ...}

1: Gilbert Downgraded to Tropical Storm; Floods And Tornadoes Feared
Tags: {Weather=0.116, Caribbean=0.044, Regional=0.033, News=0.048}
2: BROWNSVILLE, Texas (AP)
Tags: {Media=0.056, Analysis=0.047, Breaking=0.047, News=0.031}
3: Hurricane Gilbert weakened to a tropical storm as it blustered northwestward ...
Tags: {Weather=0.069, Caribbean=0.045, Regional=0.033, News=0.027}
4: At least 109 people have been killed and more than 800,000 left homeless since ...
Tags: {Caribbean=0.07, Regional=0.054}
5: One person died today in Texas after being injured in a tornado.
Tags: {Weather=0.13, Breaking=0.101, News=0.065}
6: But by this morning, the once massive storm's maximum sustained winds had ...
Tags: {Caribbean=0.088, Regional=0.062}

شکل 3- مثالی از گروه‌های ایجاد شده به وسیله DMOZ

تمامی جملاتی که در 20 وب سایتی که با استفاده از موتور جستجوی یاهو یافته شده است را درنظر گرفته و روی درخت DMOZ که گروه‌بندی موضوعات است، برای هر گره میزان tf-idf را محاسبه می‌نماید. برای این‌کار، برچسب نام هر گره و گره پدرش را در نظر گرفته و پرسشی بر حسب اجتماع این دو در یاهو انجام داده و وب سایت‌هایی که امتیاز بالاتری دارند را در نظر گرفته و پس از حذف برچسب‌های html و ریشه‌یابی به وسیله الگوریتم پورتر [9] و حذف کلمه‌های توقف ، tf-idf را در مورد آن‌ها محاسبه می‌نماییم. در معادله ذیل wi(t) برابر است با tf-idf کلمه t در طبقه فرزند i. وزن هر گره‌ای برابر است با جمع وزن فرزندانش است.

$$\acute{w}_c(t) = w_c(t) + \sum_{i \epsilon children(c)} w_i(t)$$

نحوه نگاشت جمله به درخت به این صورت می‌باشد که زیر درخت‌هایی که جمله را به صورت بهتری نمایش می‌دهند انتخاب می‌شوند. اگر یک جمله به چندین زیر درخت نگاشت شود تمامی گره‌های موجود در زیر درخت را درنظر می‌گیریم. الگوریتم ابتدا از ریشه شروع می‌شود و ضرب نقطه‌ای بین بردار جمله و تمام گره‌های بچه را مقایسه می‌کند. (ضرب نقطه‌ای بردارها میزان شباهت را بیان می‌دارد) و بعد میانگین و واریانس نتیجه تشابه را مقایسه می‌کند و محاسبه می‌کند کدام جملات درجه تشابه‌اشان از $\mu + \alpha^* \sigma$ بیشتر است. اگر ماکزیمم میزان شباهت به یک گره فرزند کمتر از میزان شباهت به گره جاری

بود، الگوریتم متوقف می‌شود. همپوشانی زیر درخت‌ها برای هر جمله محاسبه می‌شود و تعداد زیر درخت‌ها به عنوان معیاری برای مقدار اطلاعات جمله می‌باشد. اگر جمله‌ای در گره برگ طبقه بندی شود، اطلاعات خاص بیشتری نسبت به جمله‌ای که در سطوح بالاتر امتیازدهی شده‌است دارد.

ممکن است یک جمله به بیشتر از یک زیر درخت نگاشت شود به این خاطر از وزن اعتماد طبقه بندی کننده SVM استفاده می‌نماید:وزن برچسب t در سند d برابر است با:

$$w_d(t) = \sum_{i \in sentence(d)} conf(t,i)$$

که conf(t,i) مقدار اعتماد برچسب t در سند d می‌باشد.

حال از 4 خصوصیت ذیل در طبقه بندی کننده SVM استفاده می‌شود تا خلاصه ایجاد شود:

1) میزان همپوشانی برچسب که ضرب نقطه‌ای بردار برچسب جمله و بردار برچسب سند می‌باشد
2) عمق زیر درخت: عمق گره‌ای که به طور بهتری جمله را نمایش می‌دهد.: نشان دهنده آن است که چقدر به طور خاص درباره یک موضوع خاص جمله اطلاعات دارد.
3) تعداد زیردرخت: تعداد زیر درخت‌هایی که طبقه بندی کننده برای جمله استخراج نموده است، نشان دهنده میزان اطلاعات جمله می‌باشد
4) جمله در یک سوم اول یا دوم یا سوم متن قرار دارد.

با استفاده از خصوصیات بالا روی اسناد DUC که اسناد استانداردی است که بوسیله انسان خلاصه شده است و استفاده از طبقه بندی کننده SVM تمام جملاتی که بوسیله انسان در خلاصه آمده است به عنوان مثال‌های مثبت و بقیه منفی محسوب شده‌اند و فاز آموزش انجام شده است. از روش درستی بابی leave one out استفاده نموده است. به این صورت که در هر موضوعی تمامی جملات غیر از یک جمله جزو مجموعه آموزش می‌باشند. خلاصه بر اساس داده‌های تست ایجاد خواهد شد و جملات تست بر اساس امتیازشان رتبه بندی خواهند شد. معیار F1 در مورد این روش و روش [34] مقایسه نموده است و نتایج نشان داده که کیفیت خلاصه سازی افزایش یافته است.

8 - خلاصه‌سازی چند سنده

خلاصه‌سازی چند سنده زمانی به کار می‌آید که خلاصه‌ای از اسناد مربوط به یک موضوع درخواست شود. در حقیقت نمایش کوتاه از محتوای مجموعه‌ای از اسناد مرتبط به هم (از نظر عبارت، موضوع و یا گروه) می‌باشد. کار مهم در این نوع تولید خلاصه حذف افزونگی و تشخیص منابع و تفاوت‌های بین آن‌ها می‌باشد.

پروژه CLAIR در دانشگاه میشیگان پیشگام خلاصه‌سازی چند سنده می‌باشد. در این پروژه ارتباط زبان شناختی بین سندها همانند تضاد، تشابه تشخیص داده می‌شود. مرکز کلاستر مجموعه‌ای از کلمه‌های مهم و کلمه‌هایی که بیشتر از همه در کلاستر استفاده شده است می‌باشد. برای انتخاب متن خلاصه، جملاتی از هر یک از سندها انتخاب خواهد شد که کلاستر را بهتر توصیف نمایند. در هر حال غیر ممکن است که خلاصه ای که توسط کامپیوتر ایجاد می‌شود

9 - روش‌های ارزیابی خلاصه‌سازها

تحقیقاتی که انجام گرفته است حاکی از آن است که خلاصه‌سازی‌هایی که توسط انسان انجام می‌پذیرد وابسته به متن و خاص هر فرد می‌باشد. توانایی تولید خلاصه انسان بالا می‌باشد [58] و علاوه بر آن خلاصه‌سازی که توسط انسان ایجاد می‌شود خیلی وابسته به متن، فرد و پردازش پیچیده دارد. ارزیابی که روی خلاصه‌سازهای اتوماتیک مطرح انجام داده است حاکی از آن است که خلاصه‌ساز هایی که بر اساس اصول زبان شناختی هستند، بهترین کارایی را دارا می‌باشند [58].

در زبان انگلیسی معروفترین ابزار ارزیابی با استفاده از ROUGE [70] انجام می‌شود که برای محاسبه میزان شباهت متن خلاصه دستی و اتوماتیک از تکنیک چند وزنی استفاده می‌نماید. ROUGE به ارزیابی که انسان انجام می‌دهد بسیار نزدیک می‌باشد.

Precision و recall به عنوان معیارهای متداول در خلاصه‌سازی مطرح شده است. recall برابر است با تقسیم تعداد جملاتی که توسط سیستم درست تشخیص داده شده است بر تعداد جملاتی که توسط انسان درست تشخیص داده شده اند. Precision برابر است با تقسیم تعداد جملاتی که توسط سیستم درست تشخیص داده شده اند بر تعداد کل جملاتی که توسط سیستم برای خلاصه ایجاد شده اند[71]

$$Precision = \frac{SentenceCount(Summary_{SystemExtracted} \cap Summary_{Ideal})}{SentenceCount(Summary_{SystemExtracted})}$$

$$Recall = \frac{SentenceCount(Summary_{SystemExtracted} \cap Summary_{Ideal})}{SentenceCount(Summary_{ideal})}$$

مشخص است که هر چقدر طول خلاصه افزایش یابد، تفاوت بین متن خلاصه شده توسط انسان و متن خلاصه شده اتوماتیک افزایش می‌یابد.

معیار اندازه گیری f نیز بدین‌صورت تعریف شده است:

$$F_{measures} = 2 * precision/(recall + precision)$$

ارزیابی به دو گروه اصلی و فرعی تقسیم می‌شود. در ارزیابی اصلی، کیفیت خلاصه‌سازی از طریق مقایسه آن با خلاصه‌های ایده آل دستی انجام می‌شود) . در ارزیابی فرعی مشخص می‌شود خلاصه ایجاد شده

چقدر روی انجام کارهای دیگر تاثیر گذار بوده است. از جمله این کارها می‌توان به پرسش- پاسخ، درک مطلب و مرتبط بودن سند به موضوع خاص نام برد [72]. در صورتی که خلاصه ایجاد شده از نوع چکیده باشد ارزیابی از طریق مقایسه محتوایی با خلاصه دست ساز انسان می‌باشد(ارزیابی محتوایی). این نوع ارزیابی نیاز به تحلیل زبان شناختی دارد. مثلا فاصله بین بردار تکرار کلمات خلاصه ایجاد شده توسط انسان و سیستم محاسبه می‌شود. نوع دیگر ارزیابی، ارزیابی موضوعی می‌باشد. در این حالت از افراد خواسته می‌شود که خلاصه سیستم را از دو دیدگاه ارزیابی و امتیاز دهی نمایند. 1) خلاصه سیستم به چه میزان از محتویات مهم مقاله اصلی را می‌پوشاند و دوم آن‌که خلاصه سیستم تا چه انداز خوانا می‌باشد (امتیاز 1 تا 4 بدهند)[70].

یک ارزیابی دیگر برای چکیده‌ها بر اساس پرسش و پاسخ می‌باشد. خوانندگان پاسخ را بررسی نمایند و قضاوت نمایند که چقدر مرتبط می‌باشد که در این صورت زمان پاسخ دهی به سوال نیز به عنوان پارامتر دیگری در نظر گرفته می‌شود [71].

تشخیص کیفیت خلاصه‌سازی اتوماتیک خصوصا در نوع استخراجی کار پیچیده ای می‌باشد، به طور کلی کارایی خلاصه‌سازی از طریق مقایسه آن با روش دستی انجام خواهد شد . روشی که از ارزیابی دستی توسط انسان استفاده نموده است محدود به توانایی افراد است. در ارزیابی دستی، خلاصه ایده آل از طریق رای اکثریت و یا از طریق اجتماع و یا اشتراک انجام می‌شود. تعداد خلاصه‌سازهای فارسی متفاوت می‌باشد و از تعداد سه نفر به بالا می‌باشد.در [50] موضوعات مختلفی (کامپیوتر، هنر، ورزشی) انتخاب شده‌اند و خلاصه دستی با خلاصه منتج شده از خلاصه اتوماتیک مقایسه شده است.

در زبان انگلیسی، مجموعه خلاصه‌های تولید شده توسط انسان به نام DUC وجود دارد که می‌توانند به عنوان benchmark استفاده شود. در [73] کار خود را با خلاصه‌ساز coernicus مقایسه نموده‌اند و از مجموعه داده DUC به عنوان benchmark استفاده نموده‌اند. در [74 ,71] از مجموعه داده‌های reuter [75] استفاده شده است که شامل خلاصه‌های اخبار می‌باشد. این مجموعه شامل 1000 سند و خلاصه‌های مربوطه می‌باشد. از مجموعه داده‌های موجود در[76]که شامل 183 مقاله علمی است نیز استفاده شده است.

10 - نتیجه گیری

این مستند به بیان ادبیات خلاصه‌سازی در سراسر دنیا و چالش‌های زبان فارسی با رویکرد خلاصه‌سازی پرداخت. از آن‌جایی که خلاصه‌سای چکیده نیاز به فهم متن در سطح بالا دارد، امروزه تمرکز خلاصه‌سازی در سراسر دنیا روی خلاصه‌سازی استخراجی می‌باشد. به هر میزان که از اطلاعات زبان‌شناختی زبان‌فارسی در امر خلاصه‌سازی استفاده شود به همان میزان خلاصه دقیق‌تری به وجود خواهد آمد ولی این مساله در قبال پرداخت هزینه و تلاش بیشتری خواهد بود. در زبان فارسی تا کنون ایجاد زیر ساخت‌های مناسب برای خلاصه‌سازی دقیق متون همانند ریشه‌یاب، تشخیص دهنده کران جملات، الگوریتم‌های حذف کلمه‌های غیر مفید، تجزیه‌کننده کلامی متن و پایگاه داده لغوی در سطح جامع و کامل انجام نپذیرفته است. مجموعه استانداردی از متون و خلاصه‌های ایده‌آل آن‌ها همانند آن‌چه در زبان انگلیسی موجود می‌باشد، [65] موجود نیست که کار ارزیابی خلاصه‌سازهای فارسی را مشکل می‌نماید.

همان‌طور که اشاره شد زبان فارسی چالش‌های پیش پردازش و پردازش متن متعددی دارد. در زمینه پیش پردازش متن، تشخیص اسامی و افعال مرکب، مشخص کردن ریشه افعال و اسامی، تشخیص اختصارات، کران کلمات، مرز جملات و تطابق دادن نگارش‌های مختلف، تشخیص اسامی خاص و یکسان‌سازی متون چالش‌های متعددی وجود دارد که پرداختن به هر کدام از آن‌ها بسیار ضروری می‌باشد و هر یک نیازمند صرف تحقیق و تلاش بسیاری می‌باشد.

کارهایی که تا به کنون در زمینه خلاصه سازی فارسی انجام شده است عموما بر پایه پردازش سطحی متن انجام پذیرفته است و مناسب متون خبری می‌باشد و در زمینه پیش پردازش متون به ایجاد پایگاه داده غیر جامع از کلمه‌های توقف اکتفا شده است.

در زمینه پردازش معنایی متون کار چندانی در زبان فارسی انجام نگرفته است و عموما از پایگاه داده غیر جامع که بسیاری از کلمات و یا ارتباط بین کلمات در نظر گرفته نشده است استفاده شده به این خاطر نیاز به وجود پایگاه داده لغوی همانند WordNet برای زبان فارسی که رابطه‌های بین کلمات همانند هم معنی، متضاد، جز به کل و شمول را به طور کامل و جامع بپوشاند امری ضروری به نظر می‌رسد. این‌کار باعث می‌شود که متن خلاصه دقیق‌تری ایجاد شود. از آن‌جایی که چنین پایگاه داده لغوی موجود نمی‌باشد، امکان اعمال الگوریتم‌های مختلف مطرح در زمینه پرداز معنایی متن امکان پذیر نمی‌باشد. ایجاد چنین پایگاه داده لغوی طی دو مرحله می‌تواند انجام پذیرد:

1) ایجاد هسته مرکزی پایگاه داده که مفاهیم اصلی و پایه‌ای را بپوشاند. در این مرحله می‌توان مفاهیم اصلی که در پایگاه داده‌های لغوی دیگر زبان‌ها ایجاد شده‌است را در پایگاه داده قرار داد.
2) توسعه دادن پایگاه داده ایجاد شده در مرحله اول تا مفاهیم جزئی‌تر را پوشش دهد.

در ایجاد چنین پایگاه داده‌ای نیاز به دانستن ریشه کلمات، نقش کلمه، معنی کلمه و به طور کلی نیاز به اطلاعات زبان‌شناختی وسیعی است.

در زمینه پردازش سطح کلامی متن، نیاز به تجزیه کننده کلامی زبان فارسی می‌باشد تا جملات را به درختی تجزیه نماید که ارتباط بین جملات و بخش‌های مختلف جمله (بخش پایه و پیرو جمله) را نمایش دهد ایجاد چنین تجزیه کننده‌ای مشکل و پرهزینه خواهد بود ولی نتایج حاصل از آن به طور حتم بهتر از نتایج حاصل از پردازش سطحی متن خواهد بود که تنها به فرکانس کلمات، وجود کلمات کلیدی و خاص و مکان جمله اکتفا می‌کند. این خصوصیات منحصرا برای خلاصه‌های خبری مناسب می‌باشد.

برای ایجاد خلاصه‌سازی که متون علمی را نیز بتواند با دقت بالایی خلاصه نماید نیاز به پردازش معنایی و ساختاری متن می‌باشد که در درجه اول ایجاد پایگاه داده لغوی جامع پیش نیاز آن می‌باشد.

مراجع

[1] W.Doran, N.Stokes, and J. D. J.Carthy, “A Comapring Lexical Chain-based Summarization approches using an extractive evaluation,” in Global Word Net Conference, 2004, pp. 112-117.

[2] V. A. Yatsko, and T. N. Vishnyakov, “Some Problems in the development of Systems of Automatic Text Summarization,” *Automatic Documentation and Mathematical Linguistics*, vol. 41, no. 5, pp. 185-193, 2007.

[3] K. Spark-Jones, “Automatic Summarizing: Factors and Direction,” 1999.

[4] D. Miller, “WordNet: An On-Line Lexical Database,” *International Journal of Lexicography*, vol. 3, 1990.

[5] E. Hovy, “*Parsimonious and Profligate Approaches to the Question of Discourse Structure Relations*,” in Workshop on NLG, 1990.

[6] V. R. Vzeda, T. A. S. Pardo, and M. D. Graces, “Evaluation of Automatic Text Summarization Methods Based on Rhetorical Structure Theory,” in Eighth International Conference on Intelligent Systems Design and Applications, 2008.

[7] V. A. Yatsko, “Special Features of the Communication Syntatical Structure of Summary Utterances,” *NTI*, vol. 2, no. 11, pp. 1-5, 1993.

[8] D. I. Blumeau, and L. N. Afanasova, “Indicative Method of Computer Folding in the process of Teaching to Analytical-Synthetic Processing of Information,” 2003.

[9] M. F. Porter, “An Algorithm for Suffix Stripping,” *Program*, vol. 14, no. 3, pp. 130-137, 1980.

[10] I. Mani, and M. Maybury, “Advances in Automatic Text Summarization,” *the MIT Press*, vol. 26, no. 10, pp. 280-281, 1999.

[11] K. Ohtake, D. Kodama, and S. Masuyama, “Yet Another Summarization System with Two Modules using Emprical Knowledge,” in NTCIR Workshop, 2001.

[12] W. C. Mann, and S. A. Thompson, *Rhetorical Structure Theory: A Theory of Text Organization*, vol. ISI/RS-87-190, 1987.

[13] L. Chenghcheng, and W. Siriguleng, "Study of Automatic Text Summarization Based on Natural Language Understanding." pp. 712-714.

[14] D. Marcu, “The Theory and Practice of Discourse Parsing and Summarization.,” *MIT Press*, 2000.

[15] K. Sumita, K. Ono, T. Chino *et al.*, “A discourse structure analyzer for Japonese text.,” in In the *Proceedings of the International Conference on Fifth Generation Computer Systems*, Japan, 1992, pp. 1133-1140.

[16] T. A. S. Pardo, and M. G. V. Nunes, “Review and Evaluation of DiZer - An Automatic Discourse Analyzer for Brazilian Portuguese,” in In the Proceedings of the 7th Workshop on Computational Processing of Written and Spoken Portuguese, 2006, pp. 180-189.

[17] D. Marcu, *The Theory and Practice of Discourse Parsing and Summarization*: MIT Press, 2000.

[18] E. Villatoro-Tello, L. Villaseñor-Pineda, and M. Montes-y-Gómez, “Using Word Sequences for Text Summarization,” in TSD, 2006.

[19] T.W.Chuang, and J.Yang, “Text Summarization by Sentence Segment Extraction Using Machine Learning Algorithms”, in ACL Workshop, 2004.

[20] A. A. N. L. Freitas, and C. A. A.Kaestner, “Automatic Text Summarization using a Machine learning Approach,” in ACL, 2004.

[21] G. J. DeJong, “Fast Skimming of News Stories: The FRUMP System,” 1978.

[22] H. P. Luhn, “The Automatic Creation of Literature Abstract,” *IBM Journal of Research and Development*, pp. 159-165, 1958.

[23] S. Liu, *Theory and Methods on Text Summarization*, 2004.

[24] C. Y. Lin, and E. Hovy, "Automated Text Summarization in SUMMARIST".

[25] H. P. Edmundson, “New Methods in Automatic Abstracting,” *Journal of ACM*, pp. 264-285, 1969.

[26] M. Maybbury, “Generating Summeries from Event Data,” 1999.

[27] "GistSumm," http://www.icmc.usp.br/~taspardo/GistSumm.htm.

[28] E. Kyoomarsi, H. Khosravi, E. Eslami *et al.*, “Optimizing Text Summarization Based on Fuzzy Logic,” in Seventh IEEE/ACIS International Conference on Computer and Information Science, 2008, pp. 347-352.

[29] L. Ziheng. "Lead Detection for Improved Automatic Summarization," http://aye.comp.nus.edu.sg/meetings/050803/ziheng.pdf.

[30] M. Hassel, and N. Mazdak, “FarsiSum :A Persian Text Summarizer,” in 20th International Conference on Computational Linguistic, 2004.

[31] H. Dalianis, *SweSum- a Text Summarizer for Swedish*, Technical report, TRITANA-P0015, IPLab-174, NADA, KTH, 2000.

[32] ا. ش. شهابی, ”چکیده سازی متون فارسی ”,دومین کنفرانس بین المللی علوم شناختی, ۱۳۸۱ ,صفحه . 56.

[33] M. Y. Kan, and K. R. McKeown, “Information Extraction and Summarization: Domain Independance through Focus Types,” Colombia University, 1999.

[34] Y. Chen, X. Wang, and L. V. Yi. Guan "", pp. , 2005, “Automatic Text Summarization Based on Lexical Chains,” in Advances in Natural Computation, 2005, pp. 947-951.

[35] T. Nomoto, "Bayesian Learning in Text Summarization," in proceedings of Human Language Technology Conference on Empirical Methods in Natural Language Processing, 2005, pp. 249-256.

[36] R. Barzilay, and M. Elhadad, "Using Lexical Chains for Text Summarization," *the MIT Press*, pp. 111-121, 1999.

[37] Y.Song, K. S. Han, and H. C. Rim, "A Term weighting method based on lexical chain for automatic summarization," in CICLing, 2004.

[38] J. Morris, and G. Hirst, "Lexical cohesion computed by thesaural relations as indicator of the structure of text," *Computational linguistics*, vol. 17, no. 1, pp. 21-45, 1991.

[39] R. Barzilay, and M. Elhadad, " Lexical Chains for Text Summarization," Ben-Gurion University 1997.

[40] R. Mihalcea, "Graph-based ranking algorithms for sentence extraction, applied to text summarization," in Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, Barcelona, Spain, 2004.

[41] R. Mihalcea, "Random Walks on Text Structures," in CICLing, 2008, pp. 249-262.

[42] S. Brin, and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Comput. Netw. ISDN Syst.*, vol. 30, no. 1-7, pp. 107-117, 1998.

[43] S. Hassan, R. Mihalcea, and C. Banea, "Random-Walk Term Weighting for Improved Text Classification," in Proceedings of the International Conference on Semantic Computing, 2007.

[44] R. Mihalcea, and P. Tarau, "TextRank: Bringing Order into Texts," in EMNLP, Spain, 2004.

[45] D. Cristea, "Summarization through Discourse Structure," in CICLing 2005.

[46] J. Kupiec, J. Pedersen, and F. Chen, "Trainable Document Summarizer," in 18th ACM SIGIR, 1995, pp. 68-73.

[47] R. A. Garcia-Hernandez, and Y.Ledeneva, "Word Sequence model for single text summarization," in International Conferences in Advances in Computer-Human Interactions, 2009, pp. 44-48.

[48] Y.Ledeneva, A.Gelbukh, and R. A. Garcia-Hernandez, "Terms Derived from Frequent Sequences for Extractive Text Summarization," in CICLing, 2008, pp. 593-604.

[49] A. Zamanifar, B. Minaei-Bidgoli, and M. Sharifi, "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, 2008.

[50] H. Geng, P. Zhao, E. Chen *et al.*, "A Novel Automatic Text Summarization Study Based Term Co-Occurrence," in In the Proceeding of the 5th International IEEE Conference on Cognitive Informatics, 2006, pp. 601-606.

[51] R. A. Garci-Hernandez, R. Montiel, Y.Ledeneva *et al.*, in Internation Conference on Artificial Intelligence, 2008.

[52] Y. L. Yulia, A. Gelbukh, and H. R. García, "Keeping Maximal Frequent Sequences Facilitates Extractive Summarization,," *Research in Computing Science*, vol. 34, 2008.

[53] D. Liu, Y. He, and J. Hua, "Multi-Document Summarization Based on BE-Vector Clustering," in CICLing, 2006.

[54] H. P. Luhn, "A Statical Approach to Mechanical Encoding and Searching of Literary Information.," *IBM Journal of Research and Development*, pp. 309-317, 1975.

[55] G. Salton, and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," vol. 24, pp. 513-523, 1988.

[56] K. Ono, K. Sumita, and S. Miike, "Abstract generation based on rhetorical structure extraction," in International Conference on Computational Linguistics, 1994.

[57] D. Cristea, N. Ide, and L. Romary, "Veins Theory: a model of global discourse cohesion and coherence," in Proceedings of the 17th international conference on Computational linguistics - Volume 1, Montreal, Quebec, Canada, 1998.

[58] I. Mani, and M. T. Maybury, "Advances in Authomatic Text Summarization," *MIT Press*, pp. 442, 1998.

[59] C. N. Silla, G. L. Pappa, A. A. Freitas *et al.*, "*Automatic Text Summarization with Genetic Algorithm-Based Attribute Selection*," in LNAI 3315, 2004, pp. 305-314.

[۶۰] ز. کریمی ,م. ش. فرد" ,سیستم خلاصه سازی خودکار متون فارسی "،دوازدهمین کنفرانس بین المللی ایران, 1385.

[61] T. Chang, and W. F. Hsiao, "A hybrid approach to automatic text summarization," in CIT, 2008, pp. 65-70.

[62] H. Jing, and K. R. McKeown, "Cut and paste based text summarization," in Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, Seattle, Washington, 2000.

[63] L. Baum, "An inequality and associated maximization technique in statistical estimation of probabilistic functions of a markov process," *Inequalities*, vol. 3, pp. 1-8.

[64] M. McCord, "English Slot Grammer," *IBM*, 1990.

[65] R. Verma, P. Chen, and W. Lu, "A Semantic Free-text Summarization using Antology Knowledge," in DUC, 2007.

[66] "UMLS," http://www.nlm.nih.gov/research/umls.

[67] C. W. Wu, and L. Liu, “Ontology-based text Summarization for buisness news Articles,” in Computers and Their Application, 2003, pp. 389-392.

[68] L. Hennig, W. Umbrath, and R. Wetzker, “An Ontology-based Approach to Text Summarization,” in IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008, pp. 291-294.

[69] T. Alpcan, C. Bauckhage, and S. Agarwal, “An Efficient Ontology-Based Expert System,” in IAPR Workshop on Graph-based Representations, 2007, pp. 273-282.

[70] C. Y. Lin, and E. Hovy, “Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics”, in 2003 Language Technology Conference, Edmonton, Canada, 2003.

[71] M. Amini, and P. Gallinari, “Automatic Text Summarization Using Unsupervised and Semi-supervised Learning,” *Lecture Note on Computer Science*, vol. 2168, pp. 16-28, 2001.

[72] I. Manni, *The TIPSTER SUMMAC Text Summarization Evaluation*, The MITRE Corp., 1998.

[73] K. Bellare, A. D. Sarma, and N. Loiwal, “Generic Text Summarization using Word Net,” in International Conference on Language Resources and Evaluation, 2004.

[74] M. Amini, and P. Gallinari, “The Use of Unlabeled Data to Improve Supervised Learning for Text Summarization,” in In the Proceedings of the 25th ACM SIGIR, 2002, pp. 105-112.

[75] "Reuter (WordNews, Business News)," http://www.reuters.com/.

[76] I. Mani, G. Klein, D. House *et al.*, “SUMMAC: A Text Summarization Evaluation,” *Cambridge Journal*, vol. 8, pp. 43-68, 2002.