تحلیلی بر پیشرفت‌های چکیده‌نویسی خودکار

قاسم آزادی احمدآبادی[1]

چکیده

چکیده‌نویسی یکی از ابزارهای عمدهٔ تجزیه و تحلیل و سازماندهی مدارک است که به دو صورت دستی یا ماشینی انجام می‌شود. با پیشرفت‌هایی که در زمینهٔ نرم فزاری و سخت‌افزاری به وقوع پیوسته و در حال تکامل است توجه و گرایش به تهیه و تولید چکیده‌های خودکار گسترش پیدا کرده است. این مقاله، ضمن تقسیم‌بندی رویکردهای چکیده‌نویسی خودکار به چکیده‌نویسی تک‌سندی شامل رویکردهای سنتی، آماری (مبتنی بر مجموعه)، رویکردهای مبتنی بر ساختار گفتمان، و رویکردهای مبتنی بر دانش و نیز چکیده نویسی چندمدرکی به بررسی پیشرفت‌های صورت گرفته در این زمینه می‌پردازد. عوامل و موارد قابل ملاحظه در نظام چکیده‌سازی خودکار در سه بخش جنبه‌های دروندادی شامل ساختار متن، حوزه، سطح تخصصی، محدودیت زبان، مقیاس، رسانه، نوع، واحد و زبان؛ جنبه‌های هدف شامل موقعیت، مخاطب، و کاربرد؛ جنبه‌های بروندادی شامل محتوا، قالب، سبک، فرایند تولید، جایگزینی، و طول نیز مورد بررسی قرار گرفته است.

کلیدواژه‌ها

سازماندهی اطلاعات، چکیده‌نویسی ماشینی، چکیده‌نویسی خودکار، چکیده‌نویسی تک‌سندی، چکیده‌نویسی چندمدرکی.

مقدمه

چکیده‌نویسی زاییدهٔ عصر اطلاعات، عضو مهم خانوادهٔ اطلاع‌رسانی و ابزاری اساسی برای توسعهٔ فرهنگی و گسترش تحقیقات و پژوهش‌های نوین است. هر نوع پیشرفتی در گرو گردآوری توانایی‌ها و امکانات موجود و سپس سازماندهی آنها برای عرضهٔ مناسب به پژوهشگران، به‌منظور کمک‌رسانی به آنها در برنامه‌ریزی‌های آموزشی و پژوهشی است. چکیده‌ها برای کسانی تهیه می‌شود که از

۱. کارشناس ارشد کتابداری و اطلاع‌رسانی azadi_gh@yahoo.com

سطح خاصی از آگاهی، آموزش، و تجربه برخوردارند و با موضوع پژوهش آشنایی دارند. هدف از چکیده‌نویسی راهنمایی اینگونه محققان و صرفه‌جویی در وقت آنان در زمینهٔ گردآوری و انتخاب اطلاعات، تسریع انتقال اطلاعات، و نیز کمک به آنان در تحقیقات و تصمیم‌گیری‌هاست (۷: ۴۲– ٤٤).

در دوران کنونی با ورود ابزارها و برنامه‌های پیشرفته به حوزهٔ سازماندهی اطلاعات، حوزهٔ چکیده‌نویسی نیز دستخوش تغییرات سریع و زیادی شده و تلاش‌ها و پژوهش‌های صورت گرفته در این زمینه به ظهور سیستم‌ها و رویکردهای چکیده‌نویسی خودکار انجامیده است.

تعریف‌ها

در متون کتابداری و اطلاع‌رسانی از چکیده تعریف‌هایی ارائه شده است که به برخی از آنها اشاره می‌شود: «چکیده خلاصه‌ای است از یک نوشته که شامل فشردهٔ تمام مطالب مهم یا فشردهٔ قسمت‌های ویژه یا فهرستی از محتوای آن نوشته باشد. گاهی شامل فهرست اصطلاحات و کلیدواژه‌های متن است» (۳: ۱۲۷)؛ چکیدهٔ یک مدرک، بیان کوتاه و دقیق محتوای آن مدرک است، بدون تفسیر اضافی یا نقد آن و بدون توجه به اینکه چه کسی چکیده را نوشته است» (٤: ۱۱)؛ «چکیدهٔ یک مدرک عبارت است از بیان مختصر و صحیح یک مدرک که با سبکی برگرفته از سبک اصل مدرک نگاشته شده باشد» (۷: ٤۲)؛ «چکیده عبارت است از شناخت محتوایی یک مقاله یا یک مطلب و معرفی این محتوا با حداقل کلمات و عبارات ممکن» (٦: ۲۳)؛ و «چکیده خلاصهٔ کوتاهی است از یک کتاب، جزوه، مقاله یا رساله که نکات اصلی آن را ذکر کرده باشد» (۱: ۱۵).

چکیده‌نویسی شامل فعالیت‌هایی زنجیره‌ای است که طی آن اطلاعات موجود در مدرک در ابعاد کوچک‌تری ارائه می‌شوند. برای رسیدن به این هدف انجام دو فرایند ضروری است:

الف) پیرایش متن یا حذف اضافات: یعنی حذف نکات غیرضروری و حفظ نکات ضروری؛ و

ب) عصاره‌گیری متن: یعنی رسیدن به اصل مفهوم محتوای مدرک.

انجام دو فرایند فوق طی ٤ مرحلهٔ زیر صورت می‌پذیرد:

۱. **خواندن و درک:** اولین و مهم‌ترین وظیفهٔ چکیده‌نویسی خواندن و درک مفاهیم موجود در مدرک است. به این منظور باید چکیده‌نویس در زمینهٔ موضوع متن دارای حداقل دانش موضوعی باشد تا بتواند در خاتمه به تفسیری صحیحی از مدرک دست یابد.

۲. **انتخاب:** شامل حذف هدفمند بخش‌هایی از متن که به‌عنوان حشو و اضافات مطرح هستند. در پایان این مرحله اطلاعات مرتبط و ضروری باقی می‌ماند. در این مرحله حفظ مفهوم و ارتباط مطالب باقی مانده مهم است.

۳. **تفسیر:** با توجه به اهداف مدرک اصلی باید تفسیر گزینشی هدفمندی انجام گیرد. در این مرحله چکیده‌نویس به استقرا، قیاس،

استدلال، و استنتاج می‌پردازد.

٤. **توصیف ترکیبی- تحلیلی**: در این مرحله باید مواردی چون انسجام، هماهنگی، نظم‌پذیری، حفظ ساختار، طرح کلی معنایی و ساختار معانی بیان متن درنظر گرفته شده و چکیده‌نویس براساس نوع چکیده، سطح توصیف را که براساس تحلیل‌های قبلی مشخص شده مد نظر قرار دهد (**۲**: ٦٧).

چکیده‌نویسی خودکار چیست؟

مبنای چکیده‌نویسی خودکار این است که اگر براساس معیار تکرار می‌توان نرم‌افزار را برمبنای گزینش اصطلاحات از مدارک طراحی کرد، پس باید بتوان آن را به نحوی برنامه‌ریزی کرد که جملات را نیز از مدارک انتخاب کند.

مطابق نظر مانی[2] (۲۰۰۱) چکیده‌نویسی خودکار فرایندی است که در آن رایانه بخشی از اطلاعات منبع مثلاً یک مدرک الکترونیکی را دریافت کرده، محتوای مهم آن را گزینش کرده، و آن محتوا را به شکل فشرده به کاربر نشان می‌دهد. این فرایند شامل رایانه و نرم‌افزاری ویژه برای اجرای چکیده‌سازی واقعی است. بخشی از نرم‌افزار سیستم «خلاصه‌ساز» نامیده می‌شود. خلاصه‌ساز سیستمی است که نسخهٔ فشرده شده‌ای از محتوای سند را به شکل قابل خواندن برای کاربران تولید می‌کند. یک خلاصه‌ساز خوب باید تنها اطلاعات مهم محتوا را انتخاب و استخراج کند (**۲۸**: ٦٢).

چکیدهٔ خودکار با استفاده از روش‌های علمی و تجربی مبتنی بر فناوری اطلاعات و رایانه تهیه می‌شود. در این شیوه چهار اصل مهم باید رعایت شود:

۱. هر پارگراف از متن مدرک اصلی در یک جمله خلاصه شود؛

۲. اولین جملهٔ هر پارگراف از متن، که معمولاً بیشترین مقدار اطلاعات را دارد، استخراج شود؛

۳. جمله‌های مربوط به موقعیت‌هایی از متن مدرک که با حروف بزرگ نوشته شده‌اند، استخراج گردند؛ و

٤. جمله‌هایی از متن مدرک استخراج گردند که دارای کلیدواژه‌های معینی بوده و یا دارای واژگانی پربسامد است.

این چکیده معمولاً براساس الگویی از پیش تعیین شده توسط رایانه و با استفاده از نرم‌افزارهای مخصوص تهیه می‌شود. برنامه یا نرم افزار مورد استفاده شامل دو قسمت است:

الف) اصول صوری از پیش تعیین شده مبتنی بر روش‌های علمی و تجربی فناوری اطلاعات و سبک‌سنجی رایانه‌ای؛ و

ب) پردازش زبان طبیعی که یکی از مباحث مهم و نوین در داده‌پردازی یا دانش رایانه است. در پردازش زبان طبیعی رایانه یا برنامه می‌تواند معنای کلمات و نقش دستوری آنها و ساختارهای دستوری را تشخیص دهد و مثلاً یک جمله را به یک عبارت و یک پارگراف را به یک جمله خلاصه کند (**۵**: ۱٦۱– ۱٦۲).

پردازش زبان، با ورود فناوری رایانه‌ای به زندگی انسان مورد توجه قرار گرفته است.

2. Mani

منظور از پردازش زبان طبیعی این است که سیستم قادر باشد زبان انسان را بفهمد، تحلیل کند، و حتی زبان طبیعی تولید کند. در پردازش زبان طبیعی هدف اصلی، خلق نظریه‌هایی محاسباتی از زبان با استفاده از الگوریتم‌ها و ساختارهای داده‌ای موجود در علوم رایانه است. کاربردهای پردازش زبان طبیعی به دو دستهٔ کلی تقسیم می‌شود: نوشتاری و گفتاری. کاربردهای نوشتاری آن مانند استخراج اطلاعاتی خاص از یک متن، ترجمهٔ یک متن به زبانی دیگر، و یا یافتن مستنداتی خاص در یک پایگاه دادهٔ نوشتاری (مثلاً یافتن کتاب‌های مرتبط به هم در یک کتابخانه) است. نمونه‌هایی از کاربردهای گفتاری پردازش زبان عبارت‌اند از: سیستم‌های پرسش و پاسخ انسان با رایانه، خدمات خودکار ارتباط با مشتری از طریق تلفن و سیستم‌های آموزش به دانش‌آموزان.

در کنار چکیده‌نویسی خودکار نظام استخراج خودکار متن وجود دارد. مراحل این نظام عبارت‌اند از:

۱. شناسایی بخش‌های مرتبط متن،

۲. بیرون کشیدن واحدهای کوچک از متن منبع، و

۳. کنار هم قرار دادن هر کدام از این واحدها برای تولید چکیده.

روش‌های استخراج عبارت‌اند از:

۱. روش مبتنی بر کلیدواژه: در این روش فرض بر این است که نویسنده از بعضی کلیدواژه‌ها برای بیان اندیشه اصلی خود استفاده می‌کند. اساس کار این روش بازیابی کلیدواژه‌ها از متن منبع و کنار هم گذاشتن آنهاست.

۲. روش مبتنی بر تعیین محل: این روش فرض می‌کرد که موقعیت جمله در متن می‌تواند با اهمیت آن در متن همراه باشد. برای مثال اولین و آخرین جمله‌های پاراگراف می‌تواند ایده‌های اصلی را نشان دهد و بنابراین بخشی از چکیده قرار گیرد.

۳. روش مبتنی بر عبارات راهنما و آگاهی‌رسان: این روش، داده‌های متنی را با عبارات آگاهی‌رسان یا راهنما مورد استفاده قرار می‌داد: برای مثال در متون تخصصی عباراتی نظیر: «در این مقاله...»، «این مقاله نشان می‌دهد»، «هدف از این پژوهش»، «نتایج»، و «نتیجه‌گیری»، معرف‌های خوبی برای تعیین جملات مهم خواهد بود.

٤. روش مبتنی بر متن‌کاوی: ایدهٔ این روش آن است که شاخص‌ترین کلمات جمله مهم‌تر از خود جمله است. بنابراین با شناسایی شاخص بودن کلمات ترسیم اهمیت جملات مشابه برای چکیده ممکن می‌شود. این روش بسامد واژه- مقابل بسامد جمله نامیده می‌شود (**۳۳**: ۲۱۲-۲۱۳).

رویکردهای چکیده‌نویسی خودکار

۱. چکیده‌نویسی تک‌سندی

۱-۱. رویکردهای سنتی[3]

لون[4] (۱۹۵۸) روش ساده‌ای برای ایجاد چکیده پیش‌بینی کرد. او الگوریتمی را مورد استفاده قرار داد که سند را برای تعیین برجسته‌ترین اطلاعات بررسی می‌کرد. در این الگوریتم به

3. Classical approaches

4. Luhn

هر جمله با توجه به بسامدهای آن در متن ارزش وزنی اختصاص پیدا می‌کرد.

لون به عنوان ابداع کنندهٔ چکیده‌نویسی خودکار، فرایندهای زیر را برای این منظور درنظر گرفت:

- یک سیاههٔ واژگان غیرمجاز، همهٔ کلمات غیراسمی را از پردازش‌های بعدی حذف می‌کرد،
- کلمات باقی‌مانده، شمارش شده و براساس بسامد رخداد مرتب می‌شد،
- کلماتی که بیش از تعداد معینی تکرار شده‌اند به‌عنوان کلمات مهم و پربسامد تعریف می‌شد، و
- جملاتی که این کلمات پر بسامد را دربرداشت بازیابی می‌شد.

به روش زیر برای هر جمله یک عامل مفهومی[5] تخمین زده می‌شد:

۱. تعداد خوشه‌های جملات تعیین می‌شد (خوشه، طولانی‌ترین گروه کلمات است که از کلمات مهم تشکیل شده و فاصلهٔ کلمات مهم از یکدیگر در آنها بیش از چهار کلمه نیست)،

۲. تعداد کلمات مهم در خوشه تعیین شده و مربع این رقم بر کل کلمات در خوشه تقسیم می‌شد، و

۳. عامل مفهومی به‌عنوان ارزش بالاترین خوشه یا جمع ارزش‌های همهٔ خوشه‌ها در جمله درنظر گرفته می‌شد.

براساس این الگو جملاتی که بالاترین عامل‌های مفهومی را شامل می‌شوند گزینش شده و براساس توالی رخداد آنها در متن، چکیده ایجاد می‌شد. برای کنترل تعداد جملات انتخابی نقطهٔ برشی تعریف می‌شد (**۲۷**: ۱۵۹– ۱۶۵).

ادمونسون[6] (۱۹۶۹) سه روش زیر را شناسایی و معرفی کرد:

۱. روش راهنما: این روش مشابه الگوی لون بود. براساس مجموع وزن کلمات تشکیل‌دهنده، به کلمات ارزش تعلق می‌گرفت.

۲. روش اشاره: حضور کلماتی خاص در یک جمله نشانگر این حقیقت است که این جمله احتمالاً همان جمله‌ای است که محتوای متن را نشان می‌دهد. با واژه‌نامه اشاره[7] فهرستی از کلمات دارای وزن مثبت یا منفی مشخص می‌شد. ارزش مفهومی یک جمله شامل جمع جبری وزن‌های مثبت و منفی کلمات تشکیل‌دهندهٔ آن جمله بود.

۳. روش عنوان: این روش با این فرض کار می‌کرد که کلمات موجود در عنوان‌های اصلی و فرعی، شاخص‌های خوبی برای نشان دادن محتوای مدرک هستند. جملات بر اساس تعداد کلماتی از عنوان اصلی و عنوان فرعی ارزش اعتباری دریافت می‌کردند (**۱۵**: ۲۶۴– ۲۸۵).

پولاک و زامورا[8] (۱۹۷۵) بر چکیده‌نویسی خودکار مدارک خاص تمرکز داشتند. آنها تلاش کردند که از یک الگوریتم خاصی که نتایج بهتری از یک رویکرد عمومی داشت برای چکیده‌نویسی استفاده کنند. هدف آنها

5. Significance factor
6. Edmundson
7. Cue dictionary
8. Pollock, & Zamora

توسعهٔ سیستمی بود که بروندادهای آن با استانداردهای خدمات چکیده‌نویسی شیمی[9] سازگار شود.

پولاک و زامورا الگوریتم جالبی را به‌کار بردند که به جای انتخاب، عدم پذیرش جمله را مورد استفاده قرار می‌داد. برونداد نهایی، چکیدهٔ شاخصی در حدود ۱۰ تا ۲۰ درصد حجم منبع بود. ایدهٔ اصلی این الگوریتم این بود که هر کلمه با «کدهای معنایی» مناسب بودن آن کلمه برای فشرده‌سازی را رتبه‌بندی می‌کرد. برای این منظور، پولاک و زامورا فهرستی طولانی از لغاتی تهیه کردند که بیش از ۷۰۰ اصطلاح در آن بود. در این فهرست به هر اصطلاح یک کد معنایی اختصاص یافت. کدهای معنایی نوعی سیستم شماره‌دهی بود و مشخص می‌کرد که یک کلمه یا عبارت برای آگاهی‌رسانی شاخص هست یا خیر. برای مثال عبارت «نتایج ما» کد I دریافت می‌کرد که به معنی خیلی مثبت بوده و بالاترین نمره را می‌گرفت. اگر کلمه‌ای مناسب نبود به آن کد پایین‌تری مانند B (منفی) یا حتی M (خیلی منفی) اختصاص می‌یافت. به این ترتیب هر کلمه‌ای متناسب با آن امتیاز دریافت می‌کرد و جملات بسته به امتیاز کلی آنها حذف می‌شدند (۳۴).

دو رویکرد اول راه‌حل مشابهی برای مسئله داشتند. آنها برای این کار به جملات وزن می‌دادند. سیستم لون خیلی ساده بود، زیرا فقط از بسامد واژه‌ها به‌عنوان مشخصهٔ چکیده‌نویسی استفاده می‌کرد. سیستم ادمونسون در مقایسه با سیستم لون نتایج بهتری را نشان می‌داد، زیرا ادمونسون دریافته بود که ویژگی‌هایی شبیه محل جمله و عنوان، شاخص‌های مهمی برای چکیده‌نویسی هستند. تلاش او نشان داد که استفاده از کلیدواژه‌ها به‌عنوان تنها ویژگی برای چکیده‌نویسی نتایج ضعیفی را در پی خواهد داشت.

رویکرد آخر، روش متفاوتی برای مشکل تولید چکیده‌های عمومی و تخصصی درنظر داشت. الگوریتم این روش به این صورت بود که جملاتی که کمتر اخباری و اطلاع‌دهنده محسوب می‌شدند حذف می‌شدند. این تکنیک ثابت کرد که می‌تواند مؤثرترین روش برای اسناد مرتبط با موضوع شیمی باشد.

۱-۲. رویکردهای آماری[10] (مبتنی بر مجموعه)

معمولاً اسناد یک مجموعه از انواع مختلفی هستند. رویکردهای مبتنی بر مجموعه در مرحلهٔ تجزیه و تحلیل خود با دیگر رویکردها متفاوت هستند. به این معنی که آنها به جای یک مدرک واحد کل مجموعهٔ اسناد را تجزیه و تحلیل می‌کنند. فنون یادگیری ماشینی اغلب برای این مورد استفاده قرار می‌گیرند که اطلاعات مهم در مجموعه‌ها را درک کنند. برای مثال ویژگی‌هایی نظیر محل جمله ممکن است در انواع مختلف منابع شبیه روزنامه‌ها و مقالات علمی ارزش‌های متفاوتی داشته باشد (۲۸). بنابراین، اگر الگوریتم یادگیرندهٔ مورد استفاده برای مجموعه‌ای از مقالات روزنامه‌ها تشخیص دهد که واژه در جمله‌ای در ابتدای مقاله است ارزش بیشتری پیدا

9. Chemical Abstract Service (CAS)

10. Statistical approaches

می‌کند. بعضی مواقع بسامد واژه‌ها به تنهایی کفایت نمی‌کند به همین دلیل از این الگوریتم استفاده می‌شود.

کوپیک[11] و همکارانش (۱۹۹۵) سیستم چکیده‌نویسی پیشرفته‌تری ارائه کردند. آنها از دو مجموعه استفاده کردند. یک مجموعهٔ آزمایشی و یک مجموعهٔ پرورش‌یافته. مجموعهٔ پرورش‌یافته شامل جفت‌های سند/ خلاصه بود. این خلاصه‌ها چکیده‌های تولید شده توسط متخصصان چکیده‌نویسی بود. الگوریتم آن طوری بود که احتمال مرتبط بودن جمله را محاسبه می‌کرد. کوپیک و همکارانش پنج مشخصهٔ اصلی شامل طول جمله، عبارت اشاره‌ای، موقعیت جمله در پارگراف (ابتدای پارگراف، انتهای آن، و از این قبیل)، واژه‌های موضوعی (پربسامدترین واژه‌ها)، و واژه‌های دارای حروف بزرگ (نام‌های خاص) را مورد استفاده قرار دادند. برای چکیده‌های پرورش‌یافته از فن تطبیق جمله استفاده کردند که شباهت بین جملات چکیدهٔ دستی و جملات در اصل سند را پیدا می‌کرد. بنابراین، یک جمله از چکیدهٔ دستی می‌توانست «تطبیق مستقیم» با جملهٔ دیگر در منبع باشد، «اتصال مستقیم» - به معنی دو جمله‌ای که از منبع برای تولید یک جمله در خلاصهٔ دستی استفاده شده بود - و یا «غیرقابل تطبیق» باشد(۲٤: ۵۵- ۷۰).

آون[12] و همکارنش (۱۹۹۹) از فنون مشابه کوپیک و همکاران او استفاده کردند. کار آنها فراتر از سیستم‌های عادی چکیده‌نویسی مبتنی بر بسامد بود. آنها به جای واژه‌ها از عبارات چند کلمه‌ای به‌عنوان مبنا استفاده کردند. مجموعهٔ عظیمی از مقالات روزنامه‌ها توسط آون و همکارانش پیش‌پردازش و برچسب‌زنی شد. از این مجموعه‌ها پایگاه اطلاعاتی شامل نام‌ها و عبارات چندجمله‌ای شامل نام‌ها و عبارات تولید شد و از آنجا که واژه‌ها از سرتاسر متن استخراج شده بودند، پایگاه، رکوردهای متفاوتی برای واژه‌ها و نام‌های مشابه داشت. این سیستم بعضی از اطلاعات مجموعه‌ها را یکپارچه کرده و قادر بود عبارات مجموعه را از نظر آماری جدا کند، واژه‌های خاص را با محاسبهٔ ارزش آن پیدا کرده و عبارات هم‌پیوند را شناسایی کند. با جمع‌آوری اطلاعات این مجموعه‌ها، سیستم به‌طور خودکار با حوزه‌های مختلف سازگاری پیدا می‌کرد. ویژگی دیگر این سیستم این بود که می‌توانست برای شناسایی بهتر واژه‌های خاص مورد استفاده قرار گیرد. این سیستم با معماری سرویس‌دهنده- مشتری اجرا شد. هنگامی که نام افراد از متنی که پردازش شده بود پاک می‌شد این سیستم بهتر کار می‌کرد زیرا نام افراد وزن بالایی دریافت می‌کرد، درحالی‌که برای چکیدهٔ واژه‌های مرتبط و آگاهی‌رسان نبودند (۸: ۷۱- ۸۰).

در ۱۹۹۹، هوی و لین[13]، سیستم چکیده‌نویسی را معرفی کردند که سام آریست[14] نامیده می‌شد و از شناسایی موضوع و فعالیت‌های تولیدی و تفسیری برای تولید چکیده استفاده می‌کرد. این سیستم تکنیک‌های

11. Kupiec
12. Aone
13. Hovy & Lin
14 . SUMMARIST

آماری و دانش موجود در مورد مجموعه‌ها را ترکیب کرده و چکیده و فشرده‌هایی تولید می‌کرد. دومین گام در فرایند چکیده‌نویسی در این روش، تفسیر بود. در این مرحله دو یا چند موضوع به یک مفهوم پیوند خورده بودند. این فرایند مشکل‌ترین بخش از فرایند خلاصه‌سازی بود، زیرا به دانشی نیاز داشت که به‌ندرت به‌طور واضح در متن موجود بود. آخرین مرحله در فرایند چکیده‌نویسی تولید بود. سیستم سام آریست قادر بود با بازتولید سادهٔ جملات انتخاب شده در مرحلهٔ شناسایی، محتوای چکیده‌ها را ایجاد کند. همچنین قادر به تولید فهرست‌های موضوعی با همهٔ کلیدواژه‌ها و مفاهیم هم‌پیوند بود. در نهایت، یک تولیدکنندهٔ جمله به همراه امتیازدهندهٔ جمله برای چکیده‌ها مورد استفاده قرار می‌گرفت (**۲۰**: ۸۱– ۹٤).

با افزایش حجم انتشارات الکترونیک و منابع موجود، رویکردهای مبتنی بر مجموعه‌ها متداول‌تر شده و تجزیه و تحلیل آماری به رویکردی طبیعی به شکل چکیده‌نویسی خودکار تبدیل شدند. با وجود این، هنگام تهیهٔ چکیده مشکلاتی نظیر انسجام هنوز وجود دارد و به رویکردهای پیشرفته‌تری برای تولید زبان طبیعی مورد نیاز است.

۱–۳. رویکردهای مبتنی بر ساختار گفتمان[15]

این رویکردها تلاش می‌کنند راهبردهایی را که افراد متخصص در زمینهٔ چکیده‌نویسی برای تولید چکیده استفاده می‌کنند به‌عنوان نمونه به‌کار برند. با بررسی نحوهٔ ایجاد چکیده‌های انسانی می‌توان دید بهتری نسبت به فرایند تولید چکیده پیدا کرد. از این دانش می‌توان برای ایجاد سیستم‌های بهتر چکیده‌نویسی بهره برد. چکیده‌ها معمولاً خلاصه‌های خیلی فشرده‌ای هستند که ساختار درونی سند را دنبال می‌کنند.

بوگاریو و کندی[16] (۱۹۹۷) از تجزیه و تحلیل عبارتی و روابط ارجاعی در متن استفاده کردند. این سیستم چکیده‌نویسی مبتنی بر ساختار گفتمان بود و مجموعه‌ای از عبارات و جمله‌های مهم از متن اصلی را تولید می‌کرد. معماری این سیستم شامل چندین جزء بود: پیش پردازش، تجزیه و تحلیل زبان‌شناختی، بخش‌بندی گفتمان، تجزیه و تحلیل عبارتی، جداسازی ارجاع، محاسبهٔ نمایان بودن گفتمان، و شناسایی موضوعی. این اجزا با هم کار می‌کردند، به نحوی که درونداد یکی، برونداد دیگری بود و الی آخر. این مرور فشرده به شکل فهرستی از جملات یا بخش‌هایی از جملات بود که از نظر موضوعی با یکدیگر دسته‌بندی شده بودند (۱۳).

بارزیلی و الحداد[17] (۱۹۹۹) روش بهره‌برداری از زنجیره‌های واژگانی را ارائه کردند. آنها الگوریتمی ایجاد کردند که چندین منبع اطلاعات را مورد استفاده قرار می‌داد: اصطلاحنامهٔ وردنت[18]، برچسب‌زنندهٔ گفتار[19]، تجزیه‌کنندهٔ سطحی[20]، و قسمت

15. Discourse structure based approaches
16. Boguraev & Kennedy
17. Barzilay & Elhadad
18. WordNet
19. Speech tagger
20. Shallow parser

بخش‌بندی‌کننده[21]. در این فرایند، ابتدا متن بخش‌بندی شده و سپس زنجیره‌های واژگانی تولید می‌شد. این رویه برای ساختار زنجیره‌ای، مجموعه‌ای از واژه‌ها را انتخاب کرده و سپس برای هر کدام به دنبال زنجیره مرتبط بود. برای این منظور معیار «ربط» مورد استفاده قرار گرفته و اگر زنجیرهٔ مرتبطی یافت می‌شد آن واژه مجدداً در درون زنجیره وارد می‌شد (**۹**: ۱۱۱-۱۲۱).

این افراد به این نتیجه رسیدند که زنجیرهٔ واژگانی از کلمات با بسامد پایین به علت استفاده از واژه‌هایی با بسامد بالا می‌تواند همان اطلاعات برجسته را ارائه کند. این روش برای بسیاری از کاربردهای تجاری قابل اجراست و می‌تواند چکیده‌هایی با کیفیت بالا تولید کند.

مارکو[22] (۱۹۹۵) نظریهٔ ساختار بلاغی (معانی بیانی) را برای ساختن درخت موضوعی مورد استفاده قرار داد. این الگوریتم از عبارات نشانه‌ای برای استنباط ساختار بلاغی به شکل درخت استفاده می‌کرد. گره‌ها با اسامی روابط بلاغی (برای مثال پیچیدگی، تصدیق) نامگذاری شده بود و واحدهای ابتدایی متن، برگ‌های درخت را تشکیل می‌داد. هر گره در درخت یک هسته و یک پیرو بود. فرض بر این بود که گره‌های هستهٔ اطلاعات برجسته‌تری را نسبت به گره‌های پیرو دربرداشتند (**۳۰**: ۱۲۳- ۱۳۶).

مارکو یک خلاصه‌ساز مبتنی بر گفتمان ارائه کرد که درخت‌های ساختار بلاغی را دریافت می‌کرد و آنها را برای تشکیل خلاصهٔ نهایی یک سند مورد استفاده قرار می‌داد. این کار به این علت ممکن بود که ساختار شکل‌یافتهٔ این درخت‌ها برجسته بودن عبارت‌ها برای شمارش آن را امکان‌پذیر می‌ساخت (**۲۹**).

۱-٤. رویکردهای مبتنی بر دانش[23]

تفاوت بین سیستم‌های چکیده‌نویسی مورد بحث در بالا با سیستم‌های مبتنی بر دانش این است که این روش‌ها به حوزهٔ موضوعی مشخصی مربوط بوده و حجم زیادی از اطلاعات در یک حوزهٔ خاص را پوشش می‌دهد. این ویژگی، ایجاد چکیده‌هایی از اسناد در آن حوزهٔ موضوعی را برای این سیستم‌ها ممکن می‌سازد. اشکال اصلی این سیستم‌ها که محدودیت نیز برای آنها محسوب می‌شود آن است که به‌راحتی با انواع مختلف اسناد سازگاری پیدا نمی‌کند.

لهنرت[24] (۱۹۸۱) در مورد بخش‌های طرح داستان به‌عنوان راهی برای نمایش ساختار داستان‌های روایتی نظر داده است. مبنای این ایده این است که هنگامی که افراد داستانی را می‌خوانند یک نمایش ذهنی برای آنها به‌وجود می‌آید و اطلاعاتی را منتقل می‌کنند که در داستان به وضوح به آنها اشاره نشده است. به این معنی که با استفاده از فنون سنتی چکیده‌نویسی فقط می‌توان اطلاعاتی را که به‌طور واضح در سند وجود دارد استخراج کرد.

لهنرت درک کرده بود که وقایع در یک

21. Segmentation component
22. Marcu
23 . knowledge based approaches
24. Lehnert

داستان ممکن است تأثیر مثبت، منفی، یا نامشخص بر خواننده داشته باشد. به همین علت او حالت‌های تأثیر را به‌عنوان بلوک‌های ساختمانی در نقشه پیشنهاد داد. این حالت‌ها + (مثبت)، – (منفی) یا * (حالت ذهنی نامشخص) بود (**۲۵**: ۱۷۸–۲۱٤).

لهنرت سیستم پیشنهادی خود را به‌طور کامل اجرا نکرد، اما توانست ساختاری برای چکیده‌سازی و تجزیه و تحلیل‌های سطح بالا ایجاد کند. روش او می‌تواند به‌عنوان مبنایی برای ایجاد زبان طبیعی باشد.

مک کوون[۲۵] و همکارانش (۱۹۹۵) نیز از روش مبتنی بر دانش بهره بردند. آنها دو سیستم چکیده‌نویسی استریک[۲۶] و پلان‌داک[۲۷] را تشریح کردند. اولین سیستم، تولیدکنندهٔ چکیده بود که سه جزء اصلی را مقایسه می‌کرد: امتیازدهندهٔ جمله[۲۸]، واژه‌دهنده[۲۹]، و مرورکنندهٔ جمله[۳۰]. پلان‌داک معماری متفاوتی داشت. این سیستم از طراحی گفتمان بهره برده بود و عملگرهایی نظیر صرف و حذف تکرار را مورد استفاده قرار می‌داد. در این سیستم پردازش طرح برعهدهٔ چند معیار بود: تولیدکنندهٔ اطلاعات[۳۱]، گردآورندگان واژگان، طراح گفتمان[۳۲]، هستی‌شناس[۳۳]، و تولیدکنندهٔ جمله[۳٤].

امتیازدهندهٔ جمله در سیستم استریک مجموعه‌ای از اطلاعات را به‌عنوان درونداد دریافت می‌کرد. این اطلاعات توسط تولیدکنندهٔ اطلاعات ایجاد می‌شد. بر پایهٔ این اطلاعات امتیازدهندهٔ جمله یک درخت معنایی ایجاد می‌کرد که به واژه‌دهنده منتقل می‌شد. واژه‌دهندهٔ آن درخت را پردازش کرده و آن را بر روی یک درخت نحوی واژه‌سازی شده طرح‌ریزی می‌کرد. پیش‌نویس اول شامل برونداد ترکیبی امتیازدهندهٔ جمله و معیارهای واژه‌دهنده بود. این پیش‌نویس و اطلاعات ذخیره شده به‌صورت درونداد به مرورگر جمله، که پیش‌نویس‌های نهایی را تولید می‌کرد، وارد می‌شد.

در سیستم پلان‌داک مجموعه‌ای از اطلاعات به هستی‌شناس منتقل می‌شد. این اطلاعات مجدداً توسط تولیدکنندهٔ اطلاعات ایجاد می‌شد. نقش هستی‌شناس تقویت این اطلاعات با دانش خاص حوزه و سپس

---

25. McKeown
26. STREAK
27. PLANDOC
28. Sentence scorer
29. Iexicalizer
30. Sentence reviser
31. Fact generator
32. Discourse planner

۳۳ . (ontologizer) هستی‌شناسی در ادبیات کتابداری و اطلاع‌رسانی شکل گسترش‌یافتهٔ رده‌بندی و ابزار سازماندهی منابع با هدف بازیابی متون به زبان طبیعی است. مفهوم هستی‌شناسی را می‌توان مانند اصطلاحنامه یا مجموعهٔ کلمات کنترل شده درنظر گرفت با این تفاوت که اصطلاحنامه برای انسان قابل درک است درحالی‌که هستی‌شناس برای پردازش اطلاعات توسط ماشین و انسان کاربرد دارد. با توجه به عدم تغییر در ساختار اصطلاحنامه و نیاز به ابزارهای معنایی در محیط جدید اطلاعاتی، هستی‌شناسی به‌وجود آمد تا روابط مفهومی دقیق‌تر و صحیح‌تر از روابط موجود در اصطلاحنامه‌ها را بیان کند. هستی‌شناس تشکیل شده از: لغات کنترل شده محدود و در عین حال قابل گسترش، قابلیت تغییر در رده‌ها و روابط میان اصطاحات، روابط فرعی سلسله‌مراتبی صریح و اصول منطقی برای ایجاد روابط توسعه‌یافته توسط ماشین.

34. Sentence generator

فرستادن آنها به طراح گفتمان بود. طراح گفتمان، اطلاعات غنی شده را دریافت کرده و آنها را به اطلاعات پیچیده‌تری تبدیل می‌کرد. در نهایت، مجموعه‌ای از وظایف پیچیده به واژه‌دهنده محول شده که همان کار سیستم استریک را انجام می‌داد. چکیدهٔ نهایی شامل جملاتی بود که به‌صورت خودکار از درخت‌های نحوی تولید شده بود (**۳۲**: ۷۰۲– ۷۳۳).

پلان داک، در مقایسه با استریک، رویکرد ساده‌تر و سنتی‌تری را برای تولید زبان طبیعی مورد استفاده قرار می‌داد. این دو سیستم روش‌های مناسبی برای تولید چکیده هستند.

۲. چکیده‌نویسی چندمدرکی[۳۵]

این یک رویکرد نسبتاً جدید است، اما حوزهٔ پژوهشی خیلی پرطرفداری در چکیده‌نویسی خودکار است. همان‌طور که از نام آن پیداست، تعداد مدارک مورد استفاده به‌عنوان منبع، دو یا چند مدرک است. تفاوت بین رویکردهای چندمدرکی و رویکردهای مبتنی بر مجموعه این است که مجموعه معمولاً از انواع مختلف مدارک تشکیل می‌شود، درحالی‌که در این رویکرد باید حداقل دو سند در مجموعه باشد که دارای موضوع مشابهی باشند. راه حل این مشکل استفاده از الگوریتم خوشه‌بندی برای گروه‌بندی اسناد مشابه است.

چالش‌هایی که در این حوزه وجود دارد و در روش‌های تک سندی قبلی وجود نداشت شامل موارد زیر هستند:

- **بخش زائد جمله**- حذف بخش زائد جمله هنگام پردازش تعداد زیادی از اسناد با موضوع مشابه خیلی مهم است.
- **گروه‌بندی**- برای گروه‌بندی اسناد با همدیگر از طریق موضوع به معیارهایی برای مقایسهٔ آنها با هم نیاز است.
- **ارزیابی**- افراد چکیده‌نویس به‌طور طبیعی چکیده‌هایی از اسناد چندگانه به‌وجود نمی‌آورند. بنابراین مقایسهٔ بین آنها و چکیدهٔ تولید شده به‌طور خودکار می‌تواند مسئله‌ساز باشد.

اولین تلاش در این زمینه توسط مک کوون و رادیف[۳۶] (۱۹۹۵) انجام شد. آنها سیستمی به نام سامونز[۳۷] ارائه کردند که چکیده‌سازی مقالات خبری را انجام می‌داد. هدف این سیستم تولید چکیده‌هایی روان و با طول‌های مختلف بود. سامونز مبتنی بر معماری تولید زبان طبیعی سنتی بود و دو شاخص اصلی برای طراحی محتوا و کارکردهای زبان‌شناختی داشت. طراح محتوا شامل طراح پارگراف و ترکیب‌کننده بود. بخش زبان‌شناختی از انتخابگر واژگانی[۳۸]، هستی‌شناس[۳۹] و تولیدکنندهٔ جمله[۴۰] تشکیل شده بود.

نقش طراح متن تعیین اطلاعات لازم برای چکیده بود. مجموعه‌ای از عملگرهای طراحی پیش‌بینی شده بود و توسط طراح محتوا مورد استفاده قرار می‌گرفت. این عملگرها شامل:

35. Multi-document abstracting
36. Radev
37. SUMMONS
38. Lexical chooser
39. Ontologizer
40. Sentence generator

تغییر جنبه، رد، اضافه، پالایش، و پذیرش[41] بود. هر کدام از اینها یک قانون نوشته شدهٔ دستی داشت که دو الگو را به هم ربط می‌داد و نتیجهٔ آن الگوی سومی بود که تولید شده بود. برای مثال اگر دو مقالهٔ خبری در تعارض با یکدیگر بود، یک عملگر متناقض در رابطه با الگوهای مطابق آنها مورد استفاده قرار می‌گرفت. بنابراین، یک الگوی سومی ایجاد می‌شد که با دو الگوی دیگر متفاوت بود (**۳۱**: ۳۸۱– ۳۸۹).

در مجموع، کل اجزای زبان‌شناختی سیستم پلان‌داک مجدداً مورد استفاده قرار می‌گرفت که شامل قوانین دستوری بود و ملزم شده بود که واژه‌ها را برای تولید زبان طبیعی مورد استفاده قرار دهد (**۳۲**: ۷۱۱). انتخابگر واژگانی ساختار هر جمله را با انتخاب واژه‌های مناسب برای هر نقش معنایی مدیریت می‌کرد. در نهایت، تولیدکنندهٔ جمله، جملات زبان طبیعی را با هماهنگ‌سازی آن ساختار محتوایی ایجاد می‌کرد. الگوریتم تعریف شده توسط مک کوون و رادیف چندین مرحله داشت: پردازش، ترکیب، طراحی گفتمان، تغییر شکل. الگوهای اولیه با نظم زمانی مرتب شده بودند. سپس این الگوها با استفاده از عملگرهای طراحی (تعارض، بهبود، و مانند آن) ترکیب شده و الگوی جدیدی تولید کرده که به ترتیب اولویت مرتب شده بود. در نهایت، تولیدکنندهٔ جمله پارگراف خلاصه‌ای ایجاد می‌کرد که طول متغیری داشت (**۳۱**: ۳۸٤).

در تلاشی دیگر، رادیف و همکارانش (۲۰۰۳) رویکرد جدیدی را به‌منظور چکیده‌نویسی چندمدرکی پیشنهاد دادند. این روش، چکیده‌نویسی مبتنی بر مرکز ثقل نام داشت. این رویکرد از خوشه‌هایی که با گروه‌بندی اسناد مشابه با هم ایجاد شده بود استفاده می‌کرد. اگر بردار بالاترین ارزش یا وزن یک سند به بردار مرکز ثقل آن خوشه نزدیک بود، آن سند بخشی از خوشهٔ موجود درنظر گرفته می‌شد. رادیف و همکارانش ثابت کردند که این روش می‌تواند برای چکیده‌نویسی چندسندی‌ها با موفقیت مورد استفاده قرار گیرد.

باید اذعان داشت که اصطلاح «خوشه‌بندی»[42] به مفهوم گروه‌بندی تصادفی مدارک با هم به‌وسیلهٔ موضوع درنظر گرفته شده بود. خوشه‌ها در خود فرایند چکیده‌نویسی نیز مورد استفاده قرار می‌گرفتند که شامل مقایسهٔ بین یک جمله و مرکز ثقل خوشه بود.

رادیف و دیگران مرکز ثقل را به‌عنوان «مجموعه‌ای از واژه‌هایی که از نظر آماری برای خوشه اسناد مهم است» تعریف کرده‌اند. اگر خوشه‌ها به‌عنوان دایره درنظر گرفته شوند، مرکز ثقل، مرکز آن خواهد بود. قابل ذکر است که یک سند اگر شباهت عمده‌ای با اسناد آن خوشه داشته باشد در آن خوشه وارد می‌شود. شباهت بین سند و خوشه با مقیاس شباهت کسینوسی محاسبه می‌شود.

هر خوشه مرکز ثقلی داشت که با فهرستی از وزن‌ها ارائه شده بود. مرکز ثقل فقط ارزش‌های بالاتر از یک حد خاص را شامل می‌شد. ارزش کلی مرکز ثقل برابر بود با جمع

41. Change of perspective, contradiction, addition, refinement and agreement

42. Clustering

این ارزش‌ها. همراه با اینها دو پارامتر دیگر شامل ارزش مکانی و همپوشانی جملهٔ اول در نظر گرفته می‌شد.

با اختصاص یک امتیاز به هر جمله، فرایند چکیده‌سازی با انتخاب جملات اول با بالاترین امتیاز ساده انجام می‌شد. کارکردهای امتیازدهی متفاوتی نیز شامل موقعیت، مرکز ثقل، و همپوشانی با جملهٔ اول مورد استفاده قرار می‌گرفت. ترکیب این ویژگی‌ها در یک سیستم امتیازدهی واحد بهترین چکیده‌ها را تولید می‌کرد (۳۵).

ساجیون و گایزاسکاس[43] (۲۰۰٤) نیز سیستمی مبتنی بر خوشه‌ها را عرضه کردند. سیستم آنها از فنون فشرده‌سازی برای تشکیل نمای مشخصی از خوشه‌های اسناد استفاده می‌کرد. این سیستم دو رویکرد را پوشش می‌داد چون برای دو کار مختلف مورد آزمایش قرار گرفته بود. دو رویکرد ایدهٔ خوشه‌بندی اسناد مشابه با هم را داشتند، اما رویکرد دوم به دانش بیشتری از روش اول نیاز داشت. مشخصه‌های به‌کار رفته در این سیستم شباهت خوشهٔ جمله، شباهت سند و جمله، و موقعیت قطعی سند بود. اسناد در صورتی وارد این خوشه می‌شد که در آنها شباهت به مرکز ثقل آن خوشه وجود داشت (۳٦).

تجزیه و تحلیل ابعاد مختلف چکیده‌سازی خودکار

فرایند چکیده‌نویسی به‌صورت سنتی به سه مرحله تقسیم می‌شود:

- تجزیه و تحلیل متن برای کسب نمایشی از متن،
- تبدیل آن به شکل خلاصه، و
- تولید برونداد مناسب برای ایجاد متن خلاصه.

اکثر پژوهش‌های اخیر در مورد چکیده‌نویسی خودکار با دو دلیل به تجزیه و تحلیل متن منبع اختصاص پیدا کرده است: نخست، خلاصه‌ای که نیاز اطلاعاتی را برآورده می‌سازد می‌تواند با زنجیره‌بندی ساده بخش‌هایی از متن منبع ساخته شود. دوم، ایجاد ساختار زبان طبیعی به ابزارهای قوی برای پردازش زبان طبیعی نیاز دارد که خارج از قابلیت‌های اکثر پژوهش‌های اخیر در این حوزه است. تجزیه و تحلیل متن منبع می‌تواند به‌عنوان مهم‌ترین عامل اثرگذار بر کیفیت چکیده باشد.

علاوه بر مراحل داخلی ایجاد چکیده، عوامل زمینه‌ای بسیاری بر فرایند چکیده‌سازی تأثیر دارد که عمدتاً مربوط به نوع و تعداد اسناد برای چکیده‌نویسی، رسانهٔ ارتباطی، شکل مورد نظر برای چکیده، مخاطب مورد نظر، و مانند آن است. چکیده‌سازی مؤثر به تجزیه و تحلیل ضمنی و مشروح عوامل متن بستگی دارد زیرا چکیده‌ها با توجه به نیاز اطلاعاتی که برآورده می‌سازند تنظیم می‌شوند. اسپارک جونز[44] سه جنبهٔ مهم چکیده را مورد توجه قرار داده است: درونداد، هدف، و برونداد (۳۸) که به شرح آنها پرداخته می‌شود:

43. Saggion & Gaizauskas

44. Sparck-Jones

۱. جنبه‌های درونداد[45]

ویژگی‌های متنی که باید چکیده شود روش ایجاد چکیده را مشخص می‌کند. جنبه‌های درونداد ذیل مربوط به چکیده‌سازی متن است:

ساختار متن[46]: علاوه بر محتوای متن، اطلاعات اسناد به روش‌های مختلف یافت می‌شود. برای مثال برچسب‌هایی که سرعنوان‌ها، فصل‌ها، بخش‌ها، فهرست‌ها، جدول‌ها، از این قبیل را مشخص می‌کند. اگر این اطلاعات به‌خوبی مرتب شده و بهره‌برداری شود می‌تواند برای تجزیه و تحلیل سند مورد استفاده قرار گیرد. برای مثال کان[47] (۲۰۰۳) از ساختار مقالات پزشکی برای ایجاد یک نمایش درختی از منبع استفاده کرد (**۲۲**). تفل و موئنز[48] (۲۰۰۲) مشخصات ساختاری مقالات علمی را برای ارزیابی مشارکت هر واحد متنی از مقاله تنظیم کردند برای اینکه چکیده‌ای از آن جنبه تعریف شده ساخته شود (**٤٠**).

حوزه[49]: سیستم‌های حساس به حوزه، تنها قادر به فراهم آوردن خلاصه‌هایی از متن هستند که به حوزه از پیش تعریف شده با قابلیت جابه‌جایی تعلق دارد. این محدودیت برای یک حوزهٔ موضوعی خاص معمولاً با این واقعیت که سیستم‌های تخصصی می‌توانند فنون فشرده‌سازی دانش قابل مشاهده در حوزه‌های کنترل را مورد استفاده قرار دهند، متعادل می‌شود، همان‌طور که در مورد سیستم چکیده‌سازی سامونز اتفاق افتاد. برعکس، سیستم‌های دارای هدف عمومی به اطلاعات حوزه‌ها که معمولاً در یک رویکرد سطحی‌تر برای تجزیه و تحلیل درونداد اسناد اتفاق می‌افتد بستگی دارد.

با وجود این، بعضی از سیستم‌های عمومی برای بهره‌برداری از اطلاعات تخصصی حوزه پیش‌بینی شده‌اند. برای مثال فراچکیده‌ساز دانشگاه کلمبیا چکیده‌های مختلف را برای انواع مختلف اسناد به‌کار می‌برد (**۱۷**). مولتی‌ژن[50] برای موارد ساده اختصاص پیدا کرده است (**۱۰**) و دِمز[51] با کتابشناسی‌ها و بقیهٔ اسناد سروکار دارد (**۳۷**).

سطح تخصصی[52]: یک متن ممکن است به عنوان متن معمولی، تخصصی و محدود در ارتباط با دانش موضوعی احتمالی خوانندگان متن منبع درنظر گرفته شود. این جنبه می‌تواند مشابه آنچه در مورد حوزه گفته شد مورد توجه قرار گیرد.

محدودیت زبان[53]: زبان درونداد می‌تواند عمومی یا محدود به یک زبان فرعی در یک حوزه یا مخاطبان باشد. ممکن است حفظ زبان فرعی در چکیده لازم باشد.

مقیاس[54]: فنون مختلف چکیده‌سازی باید با طول چکیده‌های دستی مطابقت داشته باشد. در واقع، تجزیه و تحلیل درونداد متن می‌تواند در هر مورد به‌طور متفاوتی اجرا

45. Input aspects
46. Document structure
47. Kan
48. Teufel & Moens
49. Domain
50. MULTIGEN
51. DEMS
52. Specialization level
53. Restriction of language
54. Scale

شود. برای مثال تعیین واحدهای مفهوم. در مورد مقالات خبری معمولاً جملات یا حتی بندها به‌عنوان حداقل واحدهای مفهومی مورد توجه قرار می‌گیرند، در صورتی که برای اسناد بلندتر شبیه گزارش‌ها یا کتاب‌ها، پاراگراف مقیاس مناسب‌تری به نظر می‌رسد. همچنین فنون بخش‌بندی متن درونداد در واحدهای مفهومی وجود دارد: برای متن‌های کوتاه‌تر، املا و نحو و برای متن‌های طولانی‌تر بخش‌بندی موضوعی معمول‌تر است (**۱۹**).

رسانه[55]: اگرچه تمرکز اصلی چکیده‌سازی بر منابع متنی است، خلاصه‌سازی اسناد غیرمتنی شبیه ویدئو، تصاویر یا جدول‌ها نیز در سال‌های اخیر ضروری شده است. پیچیدگی خلاصه‌سازی چندرسانه‌ای مانع توسعهٔ سیستم‌هایی با پوشش وسیع شده است. به این معنی که اغلب سیستم‌های چکیده‌سازی که می‌توانند اطلاعات چندرسانه‌ای را کنترل کنند به حوزه‌های خاص یا انواع متنی محدود شده‌اند (**۱۸**: ۲۱۵– ۲۳۹). با این حال تلاش‌هایی در جهت ادغام اطلاعات رسانه‌های مختلف انجام گرفته (**۱۱**) که پوشش وسیعی از سیستم‌های چکیده‌سازی چندرسانه‌ای را با بهره‌گیری از انواع مختلف اطلاعات سندی ممکن ساخته است.

گونه[56] **(نوع)**: بعضی از سیستم‌ها انواع خاصی از متون را مورد استفاده قرار می‌دهند نظیر ساختار هرمی مقالات روزنامه. بعضی از چکیده‌ها مستقل از نوع سند چکیده‌سازی شده‌اند، درحالی‌که بعضی دیگر به برخی از انواع اسناد اختصاص پیدا کرده‌اند مانند گزارش‌های پرستاری (**۱۶**)، مقالات پزشکی (**۲۲**)؛ خبرهای خبرگزاری (**۳۱**)؛ صورتجلسات، پست‌های الکترونیکی، و صفحات وب.

واحد[57]: این درونداد برای فرایند چکیده‌نویسی می‌تواند یک سند واحد یا سند چندگانه؛ متن ساده و نیز اطلاعات چندرسانه‌ای نظیر ویدئو یا رسانه‌های شنیداری باشد (**۳۹**).

زبان[58]: سیستم‌های چکیده‌نویسی خودکار ممکن است با بهره‌گیری از ویژگی‌های اسنادی که از نظر زبانی، ترکیبی هستند از نظر زبان دارای محدودیت نباشند (**۳۳**) یا ممکن است معماری آنها با ویژگی‌های زبان واقعی مشخص شود. به این معنا که در سیستمی که با زبان‌های مختلف سروکار دارد باید تنظیماتی صورت گیرد. پیشرفت فوق‌العاده‌ای که در این زمینه به‌وجود آمده است این است که بعضی سیستم‌های چندسندی قادر به پرداختن همزمان به اسناد به زبان‌های مختلف هستند (**۱۴**).

۲. جنبه‌های هدف

موقعیت[59]: سیستم‌های چکیده‌نویسی خودکار می‌توانند خلاصه‌سازی عمومی را انجام دهند یا به‌عنوان یک مرحلهٔ میانی برای پردازش زبان طبیعی، شبیه ماشین ترجمه،

55. Media
56. Genre
57. Unit
58. Language
59. Situation

بازیابی اطلاعات، یا جوابگویی به سؤال در سیستم‌های بزرگ‌تر ادغام شوند.

مخاطب[60]: در صورتی که خصوصیات کاربر در دسترس باشد، چکیده‌ها می‌توانند با آن نیازها و کاربران خاص تطبیق پیدا کنند. به‌طور مثال دانش قبلی کاربران در یک موضوع مشخص عامل مهمی در این زمینه است. در چکیده‌های «دورنما»[61] فرض بر این است که دانش قبلی خواننده ضعیف است، بنابراین، اطلاعات جامعی تدوین شده است، درحالی‌که «خبرهای محض»[62] چکیده‌هایی هستند که تنها جدیدترین اطلاعات در مورد یک موضوع مشخص را انتقال می‌دهند. «مُلخص‌ها»[63] نوع خاصی از این دست هستند، زیرا آنها اطلاعات نمایشی از مجموعه‌ای از اسناد مرتبط را جمع‌آوری می‌کنند.

کاربرد[64]: چکیده‌ها می‌توانند نسبت به کاربردهای تعیین شده حساس باشند: بازیابی متن منبع، پیش نمایش متن، به خاطر آوردن یک متن از پیش خوانده شده، و مرتب‌سازی.

۳. جنبه‌های برونداد

محتوا[65]: چکیده ممکن است تلاش کند همهٔ ویژگی‌های مرتبط متن منبع را نشان دهد یا ممکن است بر موارد خاص آن تمرکز کند، که با سؤال و موضوعات مشخص می‌شود. چکیده‌های کلی یا عمومی از متن برخاسته‌اند، درحالی‌که چکیده‌های متمرکز بر کاربر، ناشی از پرسش و بر تعیین نیاز اطلاعاتی کاربر متکی است نظیر پرسش یا کلمات مهم. بسته به نوع محتوایی که چکیده می‌شود، رویکردهای مختلفی مورد استفاده قرار می‌گیرد.

شکل[66]: برونداد سیستم چکیده‌نویسی می‌تواند متن بدون طرح یا در شکل خاصی باشد. قالب‌بندی می‌تواند به منظورهای مختلفی انجام شود: تبدیل به یک سبک از پیش تعیین شده (برچسب‌ها)، بهبود قابلیت خواندن (تقسیم‌بندی بخش‌ها، رنگی کردن متن) و ... .

سبک[67]: اگر چکیده موضوعات متن منبع را پوشش دهد از نوع تمام‌نما[68]ست؛ اگر پیمایش مختصری از موضوعات مطرح شده در اصل سند را پیش‌بینی کند راهنما[69]ست، اگر اطلاعات غیرموجود در متن منبع که اطلاعات موجود را کامل کند یا اطلاعات نهفته را نمایان سازد چکیدهٔ جامع[70] بوده (۴۰) یا اگر دید اضافه‌تری از متن خلاصه شده به‌وجود آورد انتقادی[71] خواهد بود.

فرایند تولید[72]: متن خلاصه اگر توسط اجزای لفظی متن نوشته شده باشد گزیده (منتخب)[73] خواهد بود، یا اگر تولید شده

60. Audience
61. Background summaries
62. Just-the-news
63. Briefings
64. Usage
65. Content
66. Format
67. Style
68. Informative
69. Indicative
70. Aggregative
71. Critical
72. Production process
73. Extract

باشد چکیده خواهد بود. اگر متن به‌خوبی شکل گرفته باشد و متصل باشد نوع برونداد خلاصهٔ موردنظر نسبتاً پیراسته است، در غیر این صورت ماهیتاً ناقص و پراکنده خواهد بود (برای مثال فهرستی از واژگان کلیدی). گزینه‌های میانی نیز وجود دارد که اغلب مرتبط به ماهیت اجزایی هستند که چکیده‌ها را تشکیل می‌دهد مانند متن‌های شبیه عنوان, پارگراف بلند، بندها و عبارات. علاوه بر آن، بعضی رویکردها کارهای ویرایشی را برای حل مشکل عدم انسجام و بخش‌های زاید در چکیده‌ها انجام می‌دهند و مانع هزینهٔ بالای مستقیم تولید زبان طبیعی می‌شوند. جینگ[74] و مک کوون راهبرد بازنویسی را برای بهبود کیفیت عمومی چکیده‌ها توسط عملگرهای ویرایشی شبیه حذف، تکمیل یا جایگزینی عناصر عبارتی مورد استفاده قرار دادند (**۲۱**).

جایگزینی[75]: چکیده‌ها می‌توانند در محل منبع به‌عنوان یک جانشین قرار گیرند یا اینکه به منبع متصل شوند (**۲۳**؛ **۲۶**) یا حتی در منبع ظاهر شوند (برای مثال با تغییر رنگ متن منبع).

طول[76]: طول موردنظر چکیده به‌طور قابل توجهی بر آگاهی‌رسانی نتیجهٔ نهایی تأثیرگذار است. این طول می‌تواند با درصد فشردگی (نسبت طول چکیده با توجه به طول متن اصلی) تعیین شود. معمولاً درصد فشردگی از ۱ تا ۳۰ درصد درجه‌بندی می‌شود. در مورد خلاصه‌سازی چندسندی‌ها، طول چکیده نمی‌تواند با این شیوه تعیین شود، زیرا چکیده همیشه به طول از پیش تعیین شده تبدیل می‌شود (**۱۲**: ۱۰۱) .

نتیجه‌گیری

چکیده‌نویسی به‌عنوان یکی از ابزارهای عمدهٔ تجزیه و تحلیل و سازماندهی مدارک مطرح است. در سال‌های اخیر تمرکز و تأکید بسیاری بر ایجاد سیستم‌ها و رویکردهای چکیده‌سازی خودکار صورت گرفته است. در هر کدام از روش‌ها و سیستم‌ها تلاش بر این بوده است که چکیده‌های حاصل شده به چکیده‌های دستی و انسانی نزدیک شود. در رویکردهای سنتی با روش لون به هر جمله با توجه به بسامدهای آن در متن ارزش وزنی اختصاص پیدا می‌کرد. رویکرد ادمونسون براساس وزن کلمات، حضور کلماتی خاص در یک جمله و تعداد کلمات عنوان اصلی و عنوان فرعی کار کرد. الگوریتم پولاک و زامورا نیز به جای انتخاب، عدم پذیرش جمله را مورد استفاده قرار می‌داد که برونداد آن چکیده در حدود ده تا بیست درصد حجم منبع بود. در رویکردهای آماری کوپیک از فن تطبیق جمله با پنج مشخصهٔ اصلی شامل طول جمله، عبارت اشاره‌ای، موقعیت جمله در پارگراف، پربسامدترین واژه‌ها، و نام‌های خاص استفاده کرد. در روش آون پایگاه اطلاعاتی شامل نام‌ها و عبارات چندجمله‌ای تولید و اطلاعات مجموعه‌ها را یکپارچه کرده و عبارات مجموعه را از نظر آماری جدا

74. Jing
75.Surrogating
76. Length

می‌کرد، واژه‌های خاص را با محاسبه ارزش آن پیدا کرده و عبارات هم پیوند را شناسایی می‌کرد. در سیستم سام‌آریست فنون آماری و دانش موجود در مورد مجموعه‌ها ترکیب شده و چکیده تولید می‌شد. در رویکردهای مبتنی بر ساختار گفتمان سیستمی توسط بوگاریو و کندی شامل اجزا: پیش پردازش، تجزیه و تحلیل زبان‌شناختی، بخش‌بندی گفتمان، تجزیه و تحلیل عبارتی، جداسازی ارجاع، محاسبهٔ نمایان بودن گفتمان، و شناسایی موضوعی اجرا شد. برازیلی و الحداد روش بهره‌برداری از زنجیره‌های واژگانی عرضه را عرضه کردند. در الگوریتم مارکو نیز از عبارات نشانه‌ای برای استنباط ساختار بلاغی به شکل درخت استفاده شد. در رویکردهای مبتنی بر دانش سیستم استریک سه جزء اصلی را مقایسه می‌کرد: امتیازدهنده جمله، واژه‌دهنده و مرورکنندهٔ جمله. در سیستم پلان‌داک پردازش طرح بر عهده چند معیار بود: تولیدکنندهٔ اطلاعات، گردآورندگان واژگان، طراح گفتمان، هستی‌شناسی، و تولیدکنندهٔ جمله. در چکیده‌نویسی چندمدرکی سیستم سامونز مبتنی بر معماری تولید زبان طبیعی سنتی بود و دو شاخص اصلی برای طراحی محتوا و کارکردهای زبان شناختی داشت. طراح محتوا شامل طراح پارگراف و ترکیب‌کننده بود. بخش زبان‌شناختی از انتخابگر واژگانی، هستی‌شناس و تولیدکنندهٔ جمله تشکیل شده بود. رادیف چکیده‌نویسی مبتنی بر مرکز ثقل را اجرا کرد. این رویکرد از خوشه‌هایی که با گروه‌بندی اسناد مشابه با هم ایجاد شده بود استفاده کرد. ساجیون و گایزاکسکاس سیستمی مبتنی بر خوشه‌ها را عرضه کردند.

در مواردی که با مدارک بسیار طولانی و بلند روبه‌رو هستیم، ضروری است که برنامه‌ای در اختیار داشته باشیم که بتواند جملات مهم را برای هر بخش از اثر انتخاب کرده و چاپ نماید. از آنجا که در تهیهٔ چکیده باید بر بخش‌های مهم و خاصی از مدرک تأکید شود، در نتیجه، باید برای طبقه یا فهرستی از کلمات خاص، وزن نیز تعیین کرد تا اطمینان حاصل شود که جملات انتخابی برای چکیده‌نویسی یک یا چند مورد از این کلمات را دربر دارند. بدیهی است که چکیده‌هایی که با این روش تهیه می‌شود به چکیده‌هایی که نیروی انسانی آنها را تهیه می‌کند شبیه نیست. از آنجا که ممکن است بعضی از جملات از اولین و برخی از آخرین پارگراف و شاید جملات دیگری از میانهٔ اثر انتخاب شوند ممکن است جملات کاملاً بی‌ارتباط با هم به نظر آیند. در حقیقت در قبال انتخاب جملاتی که در مجموع می‌تواند تصویر درستی از محتوای مدرک را ارائه دهد، این مسئله چندان مهم به نظر نمی‌آید.

منابع

۱. اعتمادی، پریچهر. «چکیده و چکیده‌نویسی». ***نشریه فنی مرکز مدارک علمی***، دورهٔ دوم، ٤ (۱۳۵۲): ۱۵.

۲. پینتو مولینا، ماریا. «الگوی روش‌شناختی برای چکیده‌نویسی مستند». ترجمهٔ علی مزینانی. ***پیام کتابخانه***، دورهٔ هفتم، ۱ (۱۳۷٦): ۵۵-٦۸.

۳. سلطانی، پوری. ***دانشنامه کتابداری و***

اطلاع‌رسانی. تهران: فرهنگ معاصر، ۹۷۳۱.

٤. صدیق‌بهزادی، ماندانا. ***اصول چکیده‌نویسی براساس استاندارد ایزو ۲۱٤-۹٦۷***. تهران: کتابخانهٔ ملی جمهوری اسلامی ایران، ۱۳۸۱.

۵. محمدی‌فر، محمدرضا. ***آشنایی با مدرک‌شناسی***. تهران: وزارت فرهنگ و ارشاد اسلامی، ۱۳۸۰.

٦. مساوات، جلال‌الدین. ***چکیده‌نویسی (خدمات اطلاعاتی)***. تهران: مرکز اسناد فرهنگی آسیا، ۱۳۵٦.

۷. نجیبی، مهدی. «آیین چکیده‌نویسی». **اطلاع‌رسانی**، دورهٔ نوزدهم، ۱ و ۲ (پاییز و زمستان ۱۳۸۲): ٤۲- ٤٤.

8. Aone, C. … [et al]. "A trainable summarizer with knowledge acquired from robust NLP techniques". In Mani, I. ; Maybury, M. T., editors. ***Advances in automatic text summarization.*** Cambridge: The MIT Press, 1999, pp.71 – 80.

9. Barzilay, R.; Elhadad, M. "Using Lexical Chains for text summarization". In Mani, I. and Maybury, M. T., editors. ***Advances in automatic text summarization.*** Cambridge: The MIT Press, 1999, pp. 111–121.

10. Barzilay, Regina; Elhadadand, Noemie; McKeown, Kathy. "Sentence ordering in multi document summarization". In Human Language Technology Conference (HLT'01). 2001. [on-line]. Available: http://www.ling.ohio-state.edu/acl08/

11. Benitez, A. B.; Chang, S.-F. "Multimedia knowledge integration, summarization and evaluation". In Proceedings of the 2002 International Workshop On Multimedia Data Mining in conjunction with the International Conference on Knowledge Discovery and Data Mining (MDM/KDD-2002). Edmonton, Alberta, 2002.

12. Boguraev, B.; Kennedy, C. "Salience-based content characterization of text documents". In Mani, I.; Maybury, M. T., editors. ***Advances in automatic text summarization.*** Cambridge: The MIT Press, 1997, pp. 99- 110.

13. Boguraev, B.; Bellamyand, Rachel; Swart, Calvin. "Summarization miniaturization: delivery of news to hand-helds". In North American of The Association for Computational Linguistics (NAACL'01). 2001. [on-line]. Available: http://www.cs.toronto.edu/~suzanne/naacl/

14. Chen, Hsin-His. "Multilingual summarization and question answering". In Workshop on Multilingual Summarization and Question Answering (COLING'2002). [on-line]. Available: http://www.informatik. uni-trier.de/~ley/db/conf/coling/coling2002.html

15. Edmundson, H. P. "New methods in automatic extracting". ***Journal of the Association for Computing Machinery,*** Vol.16, No. 2 (1969): 264 – 285.

16. Elhadad, Noemie; McKeown, Kathleen R. "Towards generating patient specific summaries of medical articles". In NAACL'01 Automatic Summarization Workshop. 2001. [on-line]. Available: http://www.cs. cmu.edu/~ref/naacl2001.html

17. Hatzivassiloglou, Vassileios… [et al]. "Simfinder: A flexible clustering tool for summarization". In NAACL'01 Automatic Summarization Workshop. 2001. [on-line]. Available: http://www.cs.cmu.edu/~ref/ naacl2001.html

18. Hauptmann, A. G.; Witbrock, M. J. "Informedia: News-on-demand multi- media information acquisition and retrieval". In Mark T. Maybury, ed. ***Intelligent multimedia information retrieval.*** Cambridge: AAAI/MIT Press, 1997, pp.215 – 239.

19. Hearst, Marti A. "Multi-paragraph segmentation of expository text". Annual Meeting of Association for Computational Linguistics, New Mexico: New Mexico stat university, 1994. [on-line]. Available: www.aclweb.org/anthology/P/P94/P94-1000.pdf

20. Hovy, E.; Lin, C. Y. "Automated text summarization in SUMMARIST". In Mani, I.; Maybury, M. T., editors. ***Advances in automatic text summarization.*** Cambridge: The MIT Press, 1999, pp. 81 - 94.

21. Jing, Hongyan; McKeown, Kathleen R. "Cut and paste based text summarization". In 1st Conference of the North American Chapter of the Association for Computational Linguistics, 2000.

22. Kan, Min-Yen. "Automatic text summarization as applied to information retrieval: using indicative and informative summaries". Ph.D. thesis, Columbia University, 2003.

23. Kan, Min-Yen; Klavans, Judith L.; McKeown, Kathleen R. "Domain- specific informative and indicative summarization for information retrieval". In workshop on text summarization in conjunction with the ACM SIGIR conference. New Orleans, 2001.

24. Kupiec, J.; Pedersen, J. O.; Chen, F. "A trainable document Summarizer". In Mani, I. and Maybury, M. T., editors. ***Advances in automatic text summarization.*** Cambridge: The MIT Press, 1995, pp. 55 – 70.

25. Lehnert, W. G. "Plot units and narrative summarization". In Mani, I.; Maybury, M. T., editors. ***Advances in automatic text summarization.*** Cambridge: The MIT Press, 1981, pp. 178– 214.

26. Leuski, Anton; Lin, Chin-Yew; Hovy, Eduard H. "iNeATS: Interactive multi-document summarization". In

Association for Computational Linguistics (ACL'03). 2003. [on-line]. Available: http://www.aclweb.org/archive/officers_new.html

27. Luhn, H. P. "The automatic creation of literature abstracts". ***IBM Journal of Research Development,*** Vol.2, No.2 (1958): 159 – 165.

28. Mani, I. ***Automatic summarization.*** Amsterdam, Philadephia: John Benjamins Publishing Company, 2001.

29. Mani, I.; Maybury, M. T., editors. ***Advances in automatic text summarization.*** Cambridge, MA.; MIT Press, 1999.

30. Marcu, D. "Discourse trees are good indicators of importance in text". In Mani, I.; Maybury, M. T., editors. ***Advances in automatic text summarization.*** Cambridge, MA.; MIT Press, 1995, pp. 123- 136.

31. McKeown, K. R.; Radev, D. R. "Generating summaries of multiple news articles". In Mani, I.; Maybury, M. T., editors. ***Advances in automatic text summarization.*** Cambridge: MIT Press, 1995, 381 – 389.

32. McKeown, K. R.; Robin, J.; Kukich, K. "Generating concise natural language summaries". ***Information Processing & Management,*** Vol. 31, No.5 (1995): 702 – 733.

33. Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. "GistSumm: a summarization tool based on a new extractive method". In Mamede, N. J... [et al] (eds). "6th workshop on Computational Processing of the Portuguese Language - Written and Spoken", no. Faro, Portugal (26- 27 Jun 2003). [on-line]. Available: http://www.icmc.usp.br/~taspardo/Publications.htm

34. Pollock, J.; Zamora, A. "Automatic abstracting research at chemical abstracts service". ***Journal of Chemical Information and Computer Sciences***, Vol. 15, No. 4 (1975).

35. Radev, D. R.; Jing, H.; Malgorzata Stys, D. T. "Centroid-based summarization of multiple documents: information processing and management". In DUC. Association for computational linguistics. Edmonton, Alberta, Canada, 2003. [on-line]. Available: http://www-nlpir.nist.gov/projects/duc/pubs.html

36. Saggion, H; Gaizauskas, R. "Multi-document summarization by cluster/ profile relevance and redundancy removal". In Document Understanding Conference (DUC), Boston, USA, 2004. [on-line]. Available: http://www-nlpir.nist.gov/projects/duc/pubs.html

37. Schiffman, Barry; Mani, Inderjeet; Concepcion, Kristian J. "Producing

biographical summaries: Combining linguistic knowledge with corpus statistics". In European the Association for Computational Linguistics (EACL'01). 2001. [on-line]. Available: http://eacl.coli.uni-saarland.de/

38. Sparck-Jones, Karen. "Factorial summary evaluation". In Workshop on text summarization in conjunction with the ACM SIGIR Conference , New Orleans, Louisiana, 2001.

39. Sundaram, H. "Segmentation, structure detection and summarization of multimedia sequences". Ph.D. thesis, Graduate School of Artsand Sciences, Columbia University, 2002.

40. Teufel, Simone; Moens, Marc. "Summarizing scientific articles – experiments with relevance and rhetorical status". ***Computational Linguistics,*** Vol. 28, No.4 (2002).

تاریخ تأیید: ۱۳۸۸/۸/۲٤