پرکاربردترین عملکردهای پردازش زبان طبیعی در حوزه علوم کتابداری و اطلاع‌رسانی

ناهید خوشیان*[1]، وحید رضا میرزائیان[2]

مطالعات دانش‌شناسی
سال هفتم، شماره ۲۳، تابستان ۹۹، ص ۱۱۷ تا ۱۵۰

تاریخ دریافت: ۹۸/۰۸/۱۷
تاریخ پذیرش: ۹۸/۱۲/۲۶

چکیده

هدف از پژوهش حاضر، بررسی پرکاربردترین کارکردهای پردازش زبان طبیعی در حوزه علوم کتابداری و اطلاع‌رسانی بوده است. پژوهش حاضر به روش تحلیل اسنادی یا کتابخانه‌ای و با مداقه و بررسی و تحلیل متون انجام شده است. یافته‌ها نشان داد که تاکنون کاربردهای مهمی از پردازش زبان طبیعی در حوزه‌های مختلف انجام شده است. در این پژوهش پرکاربردترین کارکردهای پردازش زبان طبیعی در حوزه علوم کتابداری و اطلاع‌رسانی عبارت بودند از: نمایه‌سازی خودکار، استخراج خودکار اطلاعات یا خلاصه‌سازی خودکار، بازیابی اطلاعات، بازیابی اطلاعات بین‌زبانی (نظام بازبین)، بازیابی اطلاعات موسیقیایی، رده‌بندی خودکار و سیستم‌های پرسش و پاسخ. نتایج نشان داد که پردازش زبان طبیعی، همچنان دارای قابلیت‌های خوب و مفیدی در حوزه‌های مختلف و ازجمله در رشته علوم کتابداری و اطلاع‌رسانی است که باید با برشمردن مزایا و هزینه‌ها، نسبت به ادغام پردازش زبان طبیعی در حوزه‌های موضوعی مختلف اقدام نمود.

واژه‌های کلیدی: بازیابی اطلاعات، پردازش زبان طبیعی، علوم کتابداری و اطلاع‌رسانی

۱.* دانشجوی دکتری. گروه علم اطلاعات و دانش‌شناسی گرایش بازیابی اطلاعات، دانشگاه الزهرا، تهران، ایران. کارمند شرکت سبوی دانش nkhooshian@gmail.com

۲. استادیار. گروه کاربرد فناوری در آموزش زبان. دانشگاه الزهرا، تهران، ایران. mirzaeian@alzahra.ac.ir

مقدمه

پردازش زبان طبیعی یکی از اموری است که با ورود فناوری رایانه‌ای به زندگی بشر، موردتوجه بسیاری از دانشمندان قرار گرفته است. به بیان دقیق‌تر، پردازش زبان طبیعی عبارت است از استفاده از رایانه برای پردازش زبان گفتاری و نوشتاری. پردازش زبان طبیعی یکی از شاخه‌های بااهمیت در حوزه گسترده هوش مصنوعی و زبان‌های برنامه‌نویسی هوشمند در دانش زبان‌شناسی است. منظور از پردازش زبان طبیعی این است که رایانه‌ای داشته باشیم که قادر باشد زبان انسان را تحلیل کند، بفهمد و بتواند زبان طبیعی تولید کند (برانتس[1]، ۲۰۱۵).

برخی از مهم‌ترین وظایف پردازش زبان طبیعی به شرح زیر است:

- تحلیل ارجاع: با فرض یک جمله یا بخش بزرگ‌تری از یک متن، تعیین اینکه چه کلماتی به امور واحدی ارجاع دارند، تحلیل ارجاع خوانده می‌شود. نمونه خاص این وظیفه، بررسی اَنفورا[2] یعنی مرجع‌داری است. در زبان‌شناسی به حالتی در جمله‌ها که درک معنای یک عنصر متنی با مراجعه به عناصر دیگر متن امکان‌پذیر می‌شود ارجاع می‌گویند. عمده ارجاعات را ضمایر شخصی، ملکی، اشاره‌ای و صفات ملکی تشکیل می‌دهند.
- تحلیل گفتمان: گفتمان‌شناسی یا تحلیل گفتمان، یا تحلیل کلام اصطلاحی کلی برای اطلاق به مطالعاتی است که زبان نوشتاری، گفتاری یا نشانه‌ای یا هرگونه پدیده نشانه‌شناختی را مورد تجزیه‌وتحلیل قرار می‌دهند. تحلیل گفتمان معمولاً یکی از زیرشاخه‌های علم زبان‌شناسی شناخته می‌شود.
- ترجمه ماشینی: ترجمه ماشینی زیرشاخه‌ای از محاسباتی است که عبارت است از ترجمه متنی از یک زبان طبیعی به زبانی دیگر توسط کامپیوتر. در سطح مقدماتی، ترجمه ماشینی، یک جایگزین ساده برای کلمات از زبان طبیعی به زبان دیگر است. با استفاده از تکنیک‌های زبان‌شناسی پیکره‌ای، ترجمه‌های پیچیده بیشتری قابل‌دستیابی هستند. همچنین این تکنیک‌ها،

1. Brants
2. anaphora

کنترل بهتر تفاوت‌های گونه‌شناسی در زبان، تشخیص عبارات و ترجمه اصطلاحات را به خوبی و درستی جدا کردن عبارات نامتعارف در متن مقدور می‌سازند.

- تقطیع صرفی: تقطیع کلمات به تکواژها و مشخص کردن طبقه آن‌ها، کوچک‌ترین یکای (واحد) زبان که دارای نقش دستوری و معنایی مستقل است، تکواژ یا علامت نام دارد. هجاها و طول آن‌ها نمی‌توانند برای بازشناختن تکواژها ابزار سودمندی باشند. معیار و سنجه بنیادین این است که تکواژ را نمی‌توان به یکاهای دستوری کوچک‌تر بخش کرد. دشواری این وظیفه بیشتر بر پیچیدگی صرف‌شناختی (ساختار کلمات) زبان موردبحث بستگی دارد.

- تشخیص هستارهای اسمی: تعیین اینکه در جریان یک متن، هر یک از آیتم‌ها با کدام‌یک از اسامی خاص، مثل افراد یا مکان‌ها ارتباط دارد و هر یک از این اسامی ذیل کدام گونه قرار دارد (مثلاً شخص، مکان، سازمان). لازم به ذکر است که اگرچه در برخی از زبان‌ها مثل زبان انگلیسی، حرف بزرگ در ابتدای کلمه می‌تواند به تشخیص هستار اسمی کمک کند، اما این اطلاعات نمی‌تواند در تشخیص گونه هستار اسمی کمک کند؛ و غالباً نیز نادقیق یا ناکافی است. چراکه مثلاً کلمه اول جمله نیز، با حرف بزرگ آغاز می‌شود. همچنین هستارهای اسمی، کلمات گوناگونی را در برمی‌گیرند که فقط برخی از آن‌ها با حروف بزرگ، آغاز می‌شوند. افزون بر این بسیاری از زبان‌های دیگر، مثل چینی یا عربی اصلاً از حروف بزرگ در ابتدای اسامی خاص استفاده نمی‌کنند. همچنین در برخی زبان‌ها مثل آلمانی، شروع کلمه با حرف بزرگ، برای جدا کردن انواع کلمات نیست، بلکه همه اسامی را با حرف اول بزرگ می‌نویسند.

- زایش زبان طبیعی: تبدیل اطلاعات از پایگاه داده رایانه‌ای به زبان قابل خوانش انسانی.

- فهم زبان طبیعی: تبدیل بخش‌هایی از متن به زبانی که بیشتر فرمال باشد، مثلاً در ساختار منطق متغیرها تا رایانه آسان‌تر بتواند از آن استفاده کند.

- نویسه‌خوانی نوری: بازشناسی خودکار متون موجود در تصاویر اسناد و تبدیل آن‌ها به متون قابل جستجو و ویرایش توسط رایانه

- برچسب‌گذاری اجزای کلام: تعیین اینکه هر کلمه‌ای در جمله یا بخشی از متن چه نقشی دارد. این برچسب‌گذاری بر اساس نقش آن کلمه در متن، مانند اسم، فعل، قید، صفت و

غیره صورت می‌گیرد. بعضی کلمات ممکن است یک یا چند برچسب داشته باشند. اگر یک کلمه بیش از یک برچسب داشته باشد، نیاز به ابهام‌زدایی دارد. به‌عنوان‌مثال بوک[1] در انگلیسی ممکن است اسم یا فعل باشد. یا کلمه اوت[2] حداقل پنج نقش مختلف می‌تواند بگیرد.

- تحلیل نحوی: تعیین نمودار درختی تجزیه (تحلیل نحوی) جمله. این نمودار، نمودار درختی بنیادی و منظمی است که ساختار نحوی یک زنجیره (رسته) را مطابق با دستور زبان (گرامر) با فرض مستقل از متن ارائه می‌کند. نحو زبان طبیعی مبهم است و هر جمله‌ای تحلیل‌های ممکن متکثری دارد.

- پاسخ‌گویی به پرسش‌ها: ارائه پاسخ به پرسش‌هایی با زبان‌های انسانی. برخی پرسش‌ها پاسخ درست و مشخصی دارند. مثلاً حرف اول کانادا چیست؟ اما برخی پرسش‌ها، پاسخ روشن و مشخصی ندارند، مانند معنای زندگی چیست؟ در کارهای جدید به این‌گونه پرسش‌ها و دیگر پرسش‌های پیچیده هم توجه می‌شود.

- استخراج روابط: تعیین نسبت‌ها و ارتباط‌ها میان هستارهای اسمی (مثلاً اینکه چه کسی با چه کسی ازدواج کرد).

- تعیین پایان جمله: یافتن حد پایانی جملات، یعنی جایی که هر جمله پایان می‌پذیرد. انتهای جمله‌ها معمولاً با نقطه یا دیگر علائم سجاوندی مشخص می‌شود؛ اما گاهی این علائم برای منظورهای دیگری استفاده می‌شوند. (مثلاً نقطه برای نشان دادن اختصارات نیز بکار می‌رود).

- تحلیل احساس: استخراج اطلاعات مربوط به گوینده سخنان بر اساس اطلاعات مجموعه‌ای از مدارک. از این اطلاعات برای تعیین تمایل نسبت به چیزی و به‌ویژه برای امور بازاریابی و تبلیغات استفاده می‌شود.

- تشخیص گفتار: هدف از تشخیص گفتار که در متون علمی بیشتر با نام بازشناسی گفتار شناخته شده است، طراحی و پیاده‌سازی سیستمی است که اطلاعات گفتاری را دریافت و

1. book
2. out

متن و فرمان گوینده را استخراج می‌کند. فناوری بازشناسی گفتار به رایانه‌ای که توانایی دریافت صدا را دارد (برای مثال به یک میکروفن مجهز است) این قابلیت را می‌دهد که گفتار کاربر را متوجه شود. این فناوری در تبدیل گفتار به متن و یا به‌عنوان جایگزینی برای صفحه‌کلید یا ماوس برای واردکردن دستورات مورداستفاده قرار می‌گیرد. سیستم‌های واکافت کننده گفتار، انواع مختلفی دارند. برخی قادرند گفتار پیوسته را شناسایی نمایند. برخی دیگر فقط می‌توانند گفتار گسسته (که بین کلمات سکوت) وجود دارد را شناسایی نمایند. همچنین این سیستم‌ها قادرند واژگان گفته‌شده توسط افراد مختلف و یا فقط توسط یک گوینده تشخیص دهند. به‌هرحال ایده‌آل‌ترین سیستم آن است که بتواند گفتار پیوسته غیروابسته به گوینده را در محیط نویزی شناسایی نماید. این سیستم‌ها با به‌کارگیری روش‌های مختلف طبقه‌بندی و شناسایی الگو قادر به تشخیص واژگان هستند که البته برای افزایش دقت در شناسایی از یک فرهنگ لغات نیز در انتهای سیستم استفاده می‌گردد.

- تقطیع گفتار: عبارت از این است که رایانه بتواند پس از دریافت صدای شخص، آن را به کلمات تفکیک کند. این عمل وظیفه‌ای فرعی ذیل تشخیص گفتار است.

- جداسازی مباحث: جداسازی بخش‌هایی از متن که هریک به مبحث متفاوتی اختصاص دارند.

- جداسازی کلمات: جدا کردن کلمات یک متن از یکدیگر. در زبان انگلیسی، کلمه‌ها با فاصله از یکدیگر جدا شده‌اند، اما در برخی از زبان‌ها چنین نیست، مثلاً در زبان‌های چینی و ژاپنی. جداسازی کلمات در این‌گونه زبان‌ها نیازمند دانش لغت‌شناسی و علم صرف کلمات در زبان مربوطه است.

- رفع ابهام از معنای کلمات: بسیاری از کلمات بیش از یک معنا دارند. بدین منظور لازم است مشخص کنیم که کدام معنا در یک بافت خاص، بیشتر از دیگر معانی مناسب است (نادکارنی، اُهنوماچادو و چاپمَن[1]، ۲۰۱۱).

1. Nadkarni, Ohno-Machado & Chapman

علم کتابداری و اطلاع‌رسانی نیز علمی است که به سازمان‌دهی، اشاعه و مدیریت اطلاعات می‌پردازد. در همین راستا، پردازش زبان طبیعی کاربردهای مختلفی در حیطه رشته کتابداری و اطلاع‌رسانی جهت انجام امور سازمان‌دهی و مدیریت اطلاعات دارد. در این مقاله به برخی از پرکاربردترین کاربردهای پردازش زبان طبیعی در رشته کتابداری و اطلاع‌رسانی می‌پردازیم. تأکید ما در این مقاله بیشتر روی مهم‌ترین کاربرد پردازش زبان طبیعی در رشته کتابداری یعنی بازیابی اطلاعات است. به‌عبارت‌دیگر بازیابی اطلاعات، حیطه‌ای از کتابداری و اطلاع‌رسانی است که پردازش زبان طبیعی در آن بسیار کاربرد داشته و مورداستفاده قرار می‌گیرد. کاربردهای پردازش زبان طبیعی در رشته کتابداری و اطلاع‌رسانی نیز که در این مقاله ذکر آن می‌رود، عبارت‌اند از نمایه‌سازی خودکار، استخراج خودکار اطلاعات یا خلاصه‌سازی خودکار، بازیابی اطلاعات بین‌زبانی (نظام بازبین)، بازیابی اطلاعات موسیقیایی، رده‌بندی خودکار و سیستم‌های پرسش و پاسخ.

بازیابی اطلاعات. از بسیاری از فنون پردازش زبان طبیعی در بازیابی اطلاعات، ازجمله ریشه‌سازی، تعیین مقوله دستوری (تجزیه اجزای کلام)، تشخیص مفاهیم مرکب (چندجزئی)، تجزیه مفاهیم مرکب و ابهام‌زدایی از معنی واژه استفاده می‌شود. کار اساسی بازیابی اطلاعات، بازیابی اسناد است. کارهای دیگر بازیابی اطلاعات نیز از فنون مشابهی استفاده می‌کنند، برای مثال خوشه‌بندی مدرک، پالایش، کشف پدیده‌های جدید و کشف پیوندها که می‌توان آن‌ها را با پردازش زبان طبیعی به یک روش مشابه با بازیابی اطلاعات در هم آمیخت (نادکارنی، اوهنو ماچادو، چاپمن، ۲۰۱۱).

کاربردهای پردازش زبان طبیعی در بازیابی اطلاعات از قرار زیر است:

- واژه‌های غیرمجاز: تقریباً تمام کاربردهای پردازش زبان طبیعی در بازیابی اطلاعات، پیش از پردازش اسناد و پرس‌وجوها، واژه‌های غیرمجاز را نادیده می‌گیرند. این عمل معمولاً کارکرد نظام در بازیابی اطلاعات را افزایش می‌دهد.
- ریشه‌سازی: عمل نگاشت واژه‌ها به بعضی از شکل‌های پایه (بن)، ریشه‌سازی نامیده می‌شود. این نیز یکی دیگر از کارکردهای پردازش زبان طبیعی در بازیابی اطلاعات است. دو روش عمده برای این منظور وجود دارد: ۱- ریشه‌سازی مبتنی بر زبان‌شناسی/ واژه‌نامه و

۲- ریشه‌سازی به سبک پورتر. روش اول از ریشه‌سازی بسیار دقیقی برخوردار است؛ اما مستلزم هزینه‌های بسیار بالای اجرا و پردازش و پوشش کمتر است. روش دوم دقت کمتری دارد، اما هزینه‌های اجرا و پردازش نیز کمتر است و معمولاً برای بازیابی اطلاعات کفایت می‌کند. ریشه‌سازی، چندین اصطلاح را به یک‌شکل پایه نگاشت می‌کند و سپس به‌عنوان یک اصطلاح، در مدل فضای برداری مورداستفاده قرار می‌گیرد. این بدان معنی است که ریشه‌سازی به‌طور متوسط، تشابهات را در بین اسناد، یا اسناد و پرس‌وجوها افزایش می‌دهد. این عمل منجر به افزایش در بازیافت شده، اما دقت بازیافت را تحت‌الشعاع قرار می‌دهد. ریشه‌سازی به‌ویژه وقتی‌که از ریشه‌سازی پورتر استفاده می‌شود، نمایه‌سازی را کاهش می‌دهد و به‌طورمعمول نتایج حاصله را تا حدودی بهبود می‌بخشد (ساندرسُن[1]، ۲۰۰۰).

ازجمله کارهای انجام‌شده در زمینهٔ ریشه‌یابی کلمات فارسی می‌توان به ریشه‌یاب بُن اشاره کرد. این ریشه‌یاب شبیه ریشه‌یاب پورتر عمل کرده و بر اساس قواعدی اقدام به حذف پسوندها و پیشوندها می‌کند. ریشه‌یاب دیگری که در زبان فارسی نتایج قابل قبولی ارائه کرده، ریشه‌یاب تهیه‌شده توسط (تقوا، بک‌لی و ساده[2]،۲۰۰۵) است. این الگوریتم نیز شبیه ریشه‌یاب پورتر عمل می‌کند، اما تفاوت‌هایی نیز دارد. برای مثال، الگوریتم ریشه‌یاب پورتر، به‌منظور تخمین محتوای اطلاعات، الگوی حروف صدادار و بی‌صدا را تشخیص می‌دهد؛ اما در فارسی بسیاری از حروف صدادار نوشته نمی‌شوند، بنابراین ریشه‌یاب فارسی از طول رشته برای تعریف کران پایین محتوای ریشه استفاده می‌کند که در حال حاضر، حداقل طول ریشه ۳ حرف است. این محدودیت در بعضی موارد باعث خطا می‌گردد، به‌ویژه زمانی که یک زیررشته که قسمتی از یک کلمهٔ کوتاه است به‌اشتباه به‌عنوان یک پسوند در نظر گرفته شود. تفاوت دیگر این دو الگوریتم آن است که این الگوریتم، برخلاف الگوریتم پورتر، پیشوندها را هم شناسایی می‌کند.

- تعیین مقوله دستوری: تعیین مقوله دستوری که به عمل تخصیص مقوله نحوی، به هر واژه موجود در متن اطلاق می‌شود، برخی از ابهامات در امر پردازش زبان طبیعی در بازیابی

1. Sanderson
2. Taghva,Beckley & Sadeh

اطلاعات را برطرف می‌سازد. برای مثال، تجزیه‌گر تصمیم می‌گیرد که واژه شیپس[1] در متن موجود به‌عنوان شکل جمع یا به‌عنوان فعل زمان حال سوم شخص مفرد بکار رفته است.

- عبارت‌های مرکب و آماری: عبارت‌های مرکب و آماری واحدهایی هستند که مفاهیم چندگانه‌ای را برای نمایه‌سازی فراهم می‌آورند. تکنیکی که در اسمارت[2] بکار گرفته شد، عبارت است از گردآوری زوجی از واژه‌های مجاز مجاور هم و استفاده از زوج‌هایی با بسامد بالای آستانه تعیین شده است. استفاده از این عبارت‌های مرکب و آماری نیز، یکی دیگر از تکنیک‌های پردازش زبان طبیعی در بازیابی اطلاعات است.

- جداسازی مقوله‌های مرکب: در بسیاری از زبان‌ها مانند زبان‌های هلندی، فنلاندی، آلمانی و سوئدی، واژه‌ها با به هم پیوستن سایر واژه‌ها در یک فرایند بازآور به وجود می‌آیند. توانایی در جداسازی مفاهیم مرکب در پردازش زبان طبیعی به کیفیت بازیابی اطلاعات منجر می‌گردد (هِجورلَند[3]، ۲۰۰۲).

- تشریح (تجزیه) دستوری سطحی: هدف تجزیه دستوری سطحی جداسازی واژه‌ها در یک جمله به عبارت‌های اصلی مانند عبارت‌های اسمی یا عبارت‌های ساده فعلی است؛ که این امر به‌نوبه خود موجب موفقیت پردازش زبان طبیعی در بازیابی اطلاعات می‌شود. در این زمینه، فنون متعددی آزمایش شده است. بهترین سیستم در بازیابی کار مشترک کنل[4] (۲۰۰۰) جهت تجزیه دستوری مبتنی است بر ماشین‌های بردار پشتیبان که به موفقیتی برابر با ۹۳/۴۸٪ نائل شده است.

- جفت‌های سر – تعدیل‌گر: اساس جفت‌های سر –تعدیل‌گر بر وابستگی‌هایی که در بین واژه‌ها وجود دارد، استوار است. برای مثال، یا حاصل از تجزیه عبارت – محور استاندارد یا با استفاده از یک تجزیه‌گر تابع. این تکنیک نیز موجب تسهیل در امر بازیابی اطلاعات می‌گردد.

1. Ships
2. Smart
3. Hjørland
4. CoNLL

- ابهام‌زدایی از مفهوم واژه: ابهام‌زدایی از مفهوم واژه، یکی دیگر از تکنیک‌های پردازش زبان طبیعی در امر بازیابی اطلاعات است که به تشخیص مفهوم درست یک واژه در متن اطلاق می‌شود و موجب موفقیت در امر بازیابی اطلاعات می‌گردد. (ونگ و اُراد[1]، ۲۰۰۵)

پردازش زبان طبیعی در بازیابی اطلاعات متنی در ادامه بررسی شده است:

پیچیدگی مربوط به زبان طبیعی، در بازیابی اطلاعات متنی برای تأمین نیازهای اطلاعاتی کاربر یک مسئله اساسی است. ازاین‌رو است که از فنون بازیابی اطلاعات متنی و پردازش زبان طبیعی برای آسان‌سازی توصیف محتوای مدرک و ارائه پرس‌وجوی کاربر، عموماً با هدف تطبیق توصیف‌ها و ارائه مدارک مرتبط که به نحو بهتری نیازهای اطلاعاتی کاربر را تأمین می‌کند استفاده می‌شود. بعبارت دیگر، نظام بازیابی اطلاعات متنی، وظایف زیر را در پاسخ به جستجوی کاربر انجام می‌دهد.

- نمایه‌سازی مدارک: در این مرحله از فنون پردازش زبان طبیعی برای تولید نمایه که دارای توصیف مدرک است، استفاده می‌شود. معمولاً هر مدرکی با استفاده از مجموعه‌ای از اصطلاحات که محتوای آن را به‌خوبی توصیف می‌نماید نمایه‌سازی می‌شود.

- وقتی کاربری پرس‌وجویی را تحت قاعده درمی‌آورد، نظام این پرس‌وجو را تحلیل می‌کند و در صورت لزوم به طریقی که محتوای مدرک نشان می‌دهد، ترجمه می‌نماید.

- ۳- نظام، توصیف هر مدرک را با توصیف پرس‌وجو مقایسه می‌کند و به کاربر مدارکی را نشان می‌دهد که توصیف‌های آن، نزدیک‌ترین توصیف به پرس‌وجو است. نتایج معمولاً به ترتیب ربط، یعنی برحسب سطح تشابه بین توصیف‌های مدرک و پرس‌وجو نمایانده می‌شود (اسپارک و لویس[2]، ۱۹۹۶).

نظام بازیابی اطلاعات قراردادی و نظام بازیابی هوشمند در ادامه بررسی شده است:

به‌طورمعمول، بازیابی اطلاعات قراردادی، بر بازیابی سندی تأکید می‌نماید که خیلی زیاد به رده‌بندی انسانی و به استفاده از انسان در تعیین راهبردهای جستجو وابسته است. درحالی‌که بازیابی اطلاعات هوشمند بر استخراج خودکار اطلاعات مفید تأکید نموده و موجبات تسهیل

1. Wang & Oard
2. Lewis &Spark

تعامل بین کاربر و نظام را با در دسترس قرار دادن ابزارهای دستیابی زبان طبیعی فراهم می‌آورد. وجود پایگاه‌های اطلاعاتی غیرکتابشناختی و بازیابی اطلاعات روی خط، مسائل مختلفی را ازجمله زبان بازنمون مدرک، زبان فرمان، انتخاب پایگاه اطلاعاتی، مسائل مربوط به کاربرپسند بودن و آسانی استفاده برای نظام بازیابی اطلاعات به وجود آورده است. چون تعداد پایگاه‌های اطلاعاتی غیرکتابشناختی از تعداد آثار کتابشناختی چاپی به‌مراتب بیشتر شده است، نیاز به بازیابی اطلاعات خودکار یا بازیابی اطلاعات هوشمند افزایش یافته است. بازیابی اطلاعات هوشمند از نظام‌های بازیابی اطلاعات قراردادی از نظر انعطاف‌پذیری، کاربرپسندی، پاسخ‌گویی، نمایه‌سازی و رده‌بندی خودکار متفاوت است (چن و همکاران[1]، ۱۹۹۶).

نقش پردازش زبان طبیعی در بازیابی اطلاعات هوشمند. بازیابی اطلاعات کنترل‌شده در مورد سند و پرس‌وجوها که با استفاده از ابزارهایی مانند (اصطلاح‌نامه‌ها، سرعنوان‌های موضوعی) که در پس سیستم مخفی هستند صورت می‌گیرد، به‌طورکلی به ترتیب بر اساس توصیف سند و زبان پرس‌وجوی مصنوعی ارائه می‌گردد. توصیف‌های سند به‌وسیله واژه‌های واحد یا گاهی به‌وسیله گروهی از واژه‌ها که از متون اسناد استخراج می‌شوند صورت می‌پذیرد. زبان پرس‌وجوی مصنوعی به‌وسیله اپراتورهای بولی، هم‌جواری و کوتاه‌سازی که هر سه از عملگرهای شناخته شده موجود در بازیابی اطلاعات است، پردازش می‌شود. بازیابی آثار به انطباق محاسباتی بستگی دارد. ازاین‌رو به هیچ تحلیل زبانی پالایش‌شده برای شمول در نظام بازیابی اطلاعات احتیاج نیست. باوجوداین اسناد الزاماً به شکل زبان طبیعی یا متن آزاد که بازیابی هوشمند برای استنتاج و پی بردن به توان بازیابی اطلاعات نیاز دارد، موجود و قابل‌استفاده می‌باشند (آدریانی[2]، ۲۰۰۰). به‌طورکلی برای پیاده‌سازی نظام بازیابی هوشمند، ترویج بین نظام بازیابی اطلاعات و فنون هوش مصنوعی که شامل پردازش زبان طبیعی، بازنمون دانش، استدلال و استنتاج است، طبیعی و بالقوه مفید و بااهمیت است. درنتیجه موفقیت یک نظام بازیابی اطلاعات هوشمند، بستگی به این دارد که تا چه حد، اصول

1. Chen
2. Adriani

پردازش زبان طبیعی مورداستفاده قرار می‌گیرد. بازیابی اطلاعات هوشمند عامل کلیدی در تشویق جستجوگران پردازش زبان طبیعی برای حرکت از نظام‌های کوچک‌مقیاس و داده‌های مصنوعی به نظام بزرگ‌مقیاس با استفاده از زبان انسانی است (گرافت[1]، ۱۹۹۵).

بازیابی اطلاعات بین‌زبانی. با گسترش روزافزون استفاده از اینترنت و غلبه بر محدودیت‌های فنی و شبکه‌ای که به مدد توسعه فناوری اطلاعات و ارتباطات حاصل شده است، کاربران و جستجوگران اطلاعات، دیگر تنها به منابع اطلاعاتی که به زبان آن‌ها نوشته شده اکتفا نمی‌کنند. دسترسی به همه اطلاعات مرتبط در دیگر زبان‌ها، اکنون نه آرزو بلکه حق طبیعی کاربران شناخته می‌شود. این تنوع زبانی، اگرچه در ابتدا مفید به نظر می‌رسد، اما می‌تواند مانعی برای دسترسی به اطلاعات تلقی شود؛ بنابراین امروزه بازیابی اطلاعات به فرایندهای سنتی آن خلاصه نمی‌شود، بلکه هدف‌های بزرگ‌تر (یعنی غلبه بر موانع زبانی در هنگام جستجو و بازیابی اطلاعات) نیز در این حوزه مطرح شده است. راه‌حل غلبه بر این مشکلات، بهره‌گیری از بازیابی اطلاعات بین‌زبانی است. در همین راستا، استفاده از فنون پردازش زبان طبیعی تأثیر بسزایی در کارآمدی بازیابی اطلاعات بین‌زبانی فارسی انگلیسی دارد. واژه بازبین برای بیان اختصاری بازیابی اطلاعات بین‌زبانی استفاده می‌شود. (علیزاده، ۱۳۸۳) بازیابی اطلاعات بین‌زبانی، نوعی از بازیابی اطلاعات است که در آن حداقل، دو زبان وجود دارد، زبان عبارت جستجو و زبان مجموعه مدارک. زبان عبارت جستجو را زبان‌اصلی و زبان مجموعه مدارک را زبان هدف یا مقصد می‌نامند. یک نظام بازیابی اطلاعات بین‌زبانی (بازبین)، مدرک را در زبانی که با زبان عبارت جستجو متفاوت است بازیابی می‌کند. البته کاربر نظام بازبین عبارت جستجو را به زبان بومی خویش ارائه می‌کند، اما مدارک دریافتی بر اساس زبان مجموعه مدارک خواهد بود. نظام بازبین، کار جستجوگرانی که به چند زبان تسلط دارند را ساده می‌کند و درعین‌حال جستجوگرانی را که تنها به یک‌زبان تسلط دارند قادر می‌سازد، عبارت جستجو را به زبان خود ارائه کنند و آنگاه با استفاده از دانش خود یا با بهره‌گیری، از کمک دیگران بین مدارک بازیابی شده، تمایز

1. Croft

قائل شوند. مدارکی که مربوط تشخیص داده شده‌اند سپس با استفاده از عامل انسانی یا ماشینی ترجمه و استفاده می‌شوند. یک نظام بازیابی اطلاعات بین‌زبانی دارای اجزایی است که یکی از مهم‌ترین آن‌ها قسمتی است که برای ترجمه عبارت‌های جستجو استفاده می‌شود. یافته‌های به‌دست‌آمده که از کار روی نظام بازیابی اطلاعات بین‌زبانی با استفاده از پردازش زبان طبیعی در امر بازیابی اطلاعات حاصل گردیده بیانگر نتایج زیر است:

۱- در هنگام استفاده از واژه‌نامه دوزبانه ماشین‌خوان، رویکرد ترجمه اولین برابرنهاده در مقایسه با رویکرد همه برابرنهاده‌ها باعث کارآمدی بیشتر (افزایش ضریب دقت) در نتایج بازیابی اطلاعات بین‌زبانی فارسی - انگلیسی می‌شود. جهت ارزیابی کارآمدی نظام بازبین میانگین دقت بازیافت در سطوح مختلف بازیابی شده است.

۲- پردازش مورفولوژیک واژه‌های عبارت جستجوی فارسی (که در واژه‌نامه موردبررسی وجود ندارد)، پیش از ترجمه آن‌ها در مقایسه با عدم انجام این پردازش، باعث کارآمدی بیشتر (افزایش ضریب دقت) در نتایج بازیابی اطلاعات بین‌زبانی فارسی - انگلیسی می‌شود.

۳- در هنگام ترجمه عبارت جستجوهای فارسی، شیوه ترجمه عبارتی در مقایسه با ترجمه واژه به واژه، باعث کارآمدی بیشتر (افزایش ضریب دقت در بازیافت) در نتایج بازیابی اطلاعات بین‌زبانی فارسی - انگلیسی می‌گردد.

۴- استفاده از روش آوانگاری (نوشتن واژه‌های ترجمه‌ناپذیر به الفبای زبان دیگر) اصطلاح‌های ترجمه‌ناپذیر یا خارج از واژه‌نامه فارسی، با حروف انگلیسی و بازیابی بر اساس آن، در مقایسه با حذف آن‌ها از عبارت‌های جستجو، باعث کارآمدی بیشتر بازیابی اطلاعات بین‌زبانی فارسی - انگلیسی می‌گردد. در همین راستا یک مدل پیشنهادی نظام بازبین فارسی - انگلیسی، برای ترجمه و بازیابی اطلاعات بین‌زبانی ارائه گردیده که در تحلیل این مدل باید گفت، نظام بازبین فارسی - انگلیسی، نظام خودکاری است که با استفاده از واژه‌نامه الکترونیکی و هم‌افزایی ابزارهای پردازش زبان طبیعی، پس از پردازش و ترجمه عبارت‌های جستجوی زبان‌اصلی (فارسی)، مدارک را از مجموعه زبان هدف (انگلیسی) بازیابی می‌کند. مدل پیشنهادی انعطاف‌پذیری مناسبی برای استفاده در جفت‌های زبانی دیگر نیز دارد و بر اساس آن می‌توان از زبان فارسی، به مجموعه مدارک در زبان‌های دیگر دست‌یافت. این

مدل از سویی کاربردی و از سوی دیگر، ارتباطی است. ارتباطی ازآن‌جهت که اجزای کاربردی نظام در جهت عملکرد بهینه، باید ارتباط نزدیکی داشته باشند. ویژگی بارز این مدل، استفاده از ابزار پردازش زبان طبیعی است که موجب افزایش دقت در ترجمه و کارآمدتر شدن بازبین می‌گردد. این مدل، اولین مدلی است که برای فرایند بازیابی اطلاعات بین‌زبانی فارسی مطرح شده و می‌تواند با انجام پژوهش‌های دیگر تکمیل گردد.

کارکردهای نظام بازبین. نظام بازبین پیشنهادی کارکردهای زیر را پشتیبانی می‌نماید. ترجمه عبارت جستجوی فارسی، تحلیل‌های پردازش زبان طبیعی و بازیابی مدارک در زبان انگلیسی (علیزاده، فتاحی و داورپناه، ۱۳۸۸). در نتیجه‌گیری می‌توان گفت دلایل زیادی وجود دارد که یک نظام بازبین به همان کارآمدی که بازیابی یک‌زبانه نائل می‌شود، دست پیدا نمی‌کند. اولین و مهم‌ترین دلیل وجود دو زبان در بازبین و ساختارهای متفاوت آن‌هاست. این دوگانگی موجب بسیاری از ابهام‌ها می‌گردد که بازیابی اطلاعات یک‌زبانه هرگز با آن مواجه نمی‌شود. دلیل دیگر آن است که متن و زمینه عبارت‌های جستجو در هنگام ترجمه چندان در نظر گرفته نمی‌شوند. چه‌بسا در مواردی مهجورترین برابرنهاده یک واژه، با توجه به مفهوم عبارت جستجو مناسب‌ترین انتخاب باشد. با استفاده از فنون پردازش زبان طبیعی از قبیل تحلیل مورفولوژیک، برچسب‌زنی انواع نقش دستوری و آوانگاری، مشخص شد که استفاده از ابزار زبان‌شناسی در فرایندهای بازیابی اطلاعات می‌تواند منجر به کارآمدتر شدن این نظام‌ها گردد. شناسایی الگوهای زبانی و ساختار واژه‌ها، یک نظام خودکار بازبین را قادر می‌سازد که مدارک را با دقت بیشتری بازیابی کند و رضایت‌مندی کاربر را افزایش دهد. در این راستا علاوه بر فنون نحوی بهره‌گیری از فنون پیشرفته‌تر هوش مصنوعی و استفاده از تحلیل‌های معناشناسی می‌تواند در آینده سامانه‌های بازبین را نسبت به نمونه‌های موجود، سامانه‌های موفق‌تری نشان دهد (سالتون و مک‌گیل[1]، ۱۹۸۳).

پردازش زبان طبیعی در بازیابی اطلاعات تصاویر. بازیابی اطلاعات تصاویر از سال ۱۹۷۰ یک موضوع تحقیقاتی فعال بوده است و اولین بار در این سال بازیابی تصویر مبتنی بر

1. Salton & Mc Gill

کلمه معرفی شده است. پس‌ازآن در سال ۱۹۹۰ بازیابی تصویر مبتنی بر محتوا ابداع شد. این روش به‌سرعت جایگزین روش قبلی شد و در کاربردهای مختلف بازیابی در حوزه‌های پزشکی، کتابخانه‌های دیجیتال، تنوع زیستی و ... استفاده می‌شد؛ اما در این روش از ویژگی‌های سطح پایین تصاویر استفاده می‌شد. این دو روش جزء روش‌های معمول و سنتی بازیابی تصاویر هستند؛ اما با توجه به افزایش حجم بی‌رویه تصاویر دیجیتال و نیاز به استخراج ویژگی‌های سطح بالا محققان به دنبال ابداع الگوریتم‌هایی بودند که پاسخ‌گوی این نیازها در بازیابی باشد. در این زمینه الگوریتم‌های پردازش زبان طبیعی یعنی الگوریتم‌های یادگیری ماشینی و یادگیری عمیق به‌گونه‌ای موفق عمل می‌کنند. یادگیری عمیق یکی از زیرشاخه‌های یادگیری ماشینی و یک رویکرد در حال ظهور است که به‌طور گسترده در حوزه هوش مصنوعی مانند پردازش زبان طبیعی و بینایی ماشینی استفاده شده است. شبکه‌های عصبی کانولوشن نیز به‌عنوان یکی از بهترین و مهم‌ترین الگوریتم‌های یادگیری عمیق جهت دسته‌بندی و بازیابی تصاویر است که از دقت بسیار بالایی در بازیابی تصاویر برخوردار است (سزاوار، فرسی و محمدزاده، ۱۳۹۵).

پردازش زبان طبیعی در بازیابی اطلاعات موسیقیایی. برخلاف داده‌های متنی که پردازش، جستجو و نمایه‌سازی آن‌ها آسان است، با تعداد زیادی رایانه در دسترس ایجاد می‌شوند و تکنیک‌های عرضه‌شده از سوی پردازش زبان طبیعی به بازیابی اطلاعات یا استخراج داده‌های متنی همچون طبقه‌بندی، تحلیل، خلاصه‌سازی، نمایه‌سازی، ترجمه، جستجو و بسیاری امور دیگر کمک می‌کنند، بااین‌وجود می‌توان موسیقی را نیز به‌عنوان یک زبان طبیعی قلمداد کرد و آن را به روشی مشابه متن پردازش نمود. بخش قابل‌توجهی از اصوات موسیقی، ترانه‌ها هستند. برخی از وظایف پردازش زبان طبیعی در حیطه بازیابی اطلاعات موسیقیایی و ترانه‌ها عبارت‌اند از: شناسایی زبان، کشف ساختار و مقوله‌بندی متن (باد[1]، ۲۰۰۱). بازیابی اطلاعات موسیقیایی در پی این هدف است که بازار گسترده و وسیع موسیقی جهان را در دسترس عموم مردم قرار دهد. برای رسیدن به این هدف بازنمودهای

1. Bod

مختلف اشخاص مرتبط با موسیقی (مثلاً ترانه‌سرایان، آهنگ‌سازان، نوازندگان و فروشندگان) و موضوعات مرتبط با موسیقی (مثلاً قطعات موسیقیایی، آلبوم‌ها و ویدئو کلیپ‌ها)، موردبررسی قرار می‌گیرند. بازیابی اطلاعات موسیقیایی یک حوزه پژوهشی میان‌رشته‌ای است که به استخراج، تحلیل و استفاده از اطلاعات درباره موسیقی و صوت می‌پردازد (لِرداهل و جَکِندوف[1]، ۱۹۸۳). نت‌نویسی موسیقی را نمی‌توان همانند متن به رایانه منتقل ساخت، اما به‌راحتی می‌توان به این مسئله فائق آمد. متن را می‌توان به‌آسانی به کلمات تقسیم کرد که یکی از ویژگی‌های اولیه پردازش زبان طبیعی و بازیابی اطلاعات است؛ اما این مطلب در مورد موسیقی صدق نمی‌کند. در این مورد می‌توان این‌گونه پاسخ داد که زبان‌های طبیعی همچون زبان تایلندی وجود دارند که از چیزی برای تقسیم کلمات استفاده نمی‌کنند (باد، ۲۰۱۲). بازیابی اطلاعات موسیقیایی موضوعی میان‌رشته‌ای است، زیرا از سویی شامل عناصر اساسی موسیقی و از سوی دیگر دربردارنده عناصری از علم اطلاعات است. برخی موسیقی‌ها تک‌صدایی هستند؛ یعنی، در هرلحظه تنها، صدای یک نت وجود دارد؛ اما بیشتر موسیقی‌های غربی چندصدایی هستند، یعنی در هرلحظه چندین نت به گوش می‌رسد. چندصدایی بودن موسیقی، بازیابی اطلاعات موسیقیایی را دشوارتر می‌سازد. یکی از مشکلات بازیابی در خصوص موسیقی‌های چندصدایی مسئله برجستگی است؛ یعنی مسئله تشخیص میزان اهمیت یک عنصر در موسیقی، چه این عنصر نت باشد و چه آکورد، ملودی یا هر چیز دیگر. تکنیک‌های پردازش زبان طبیعی، فراوانی در بازیابی اطلاعات بکار برده شده‌اند؛ اما نتایج چندان موفقیت‌آمیز نبوده است (اسچدل[2]، ۲۰۱۳).

پژوهش موسیقی موازی با پردازش زبان طبیعی. در ابتدا باید به بررسی این پرسش پرداخت که آیا تکنیک‌های تجزیه کردن در پردازش زبان طبیعی را می‌توان در تجزیه موسیقیایی مورداستفاده قرار داد یا خیر. مشکل اصلی در موسیقی، همانند پردازش زبان طبیعی ابهام است. چندین ساختار مختلف ممکن است با یک، سکانس موسیقیایی هم‌ساز باشند. درحالی‌که شنونده، معمولاً فقط یک ساختار را می‌شنود. بهترین تجزیه‌کننده احتمالی

1. Lerdahl & Jackendoff
2. Schedl

می‌تواند به‌درستی، ۸۵/۹ درصد از عبارات را در آزمون مجموعه‌ای از هزار ترانه محلی پیش‌بینی کند. برای در نظر گرفتن موسیقی به‌عنوان زبان طبیعی، باید نشان داد که پردازش موسیقی در همان طبقه مسائل پردازش زبان طبیعی قرار دارد. سطوح پردازش متن توسط پردازش زبان طبیعی که در جدول ۱ فهرست شده، از ضبط کردن (صدا، یا گفتار) تا فهم (معنای یک گفتار) را شامل می‌شود. این سطوح در مورد موسیقی نیز وجود دارد. موسیقی را شبیه به زبان طبیعی، می‌توان ضبط کرد و به‌عنوان یک شکل موج نشان داد. در سطح آواشناسی تلاش می‌شود که ساختار یک صدا بررسی شود و نت‌ها یا سازها از یکدیگر تفکیک شوند. باوجوداین، موسیقی در این عرصه بسیار پیچیده‌تر است و وظیفه بازشناسی صدا، با مشکلات اساسی مواجه است.

جدول ۱. سطوح پردازش زبان طبیعی یا وظایف مربوطه در پژوهش موسیقی

عرصه‌های پژوهش موسیقی	سطح NLP
تحلیل شکل موج، سیگنال‌های شنیداری	آواشناسی
شناسایی رویدادهای صوتی	واج‌شناسی
نمادهای متن موسیقی، نمادسازی	ریخت‌شناسی
مدل N-grams، تجزیه	نحو
هارمونی‌ها، سطح عبارت	معنی‌شناسی
برجسته‌سازی عبارت‌ها، صوت	کاربردشناسی
تفسیرها، بافت قطعه	گفتمان

دومین شباهت مهم از این امر حاصل می‌شود که در هر دو حوزه از نت‌نویسی نمادین استفاده می‌شود. متن‌های موسیقیایی نیز از نویسه‌هایی تشکیل می‌شوند که نت خوانده می‌شوند. شبیه به ریخت‌شناسی و نحو پردازش زبان طبیعی، موسیقی نیز ساختار دستور زبان – مانند پنهان و قواعدی پنهان دارد. بخشی از آن هارمونی است. این قواعد مشخص می‌کند که چگونه کلمات (نت‌ها) کنار یکدیگر قرار بگیرند؛ و چگونه می‌توان با استفاده از آن‌ها عبارات خوش‌فرم ساخت. نت‌ها و وابستگی آن‌ها را می‌توان به نحو موسیقی مرتبط دانست؛ اما پیشروی‌های هارمونیک در کاربردشناسی قطعه، بالاترین سطح پردازش زبان طبیعی یعنی

(گفتمان)، در موسیقی به شکل ایده‌ها، میل یا اشتیاق (موسیقی رمانتیک) سازنده اثر و نیز تصاویر و کارهایی در پشت آن (موسیقی برنامه‌ای) رواج دارد (اوراماس[1]، ۲۰۱۴؛ باد، ۲۰۰۲). تکنیک‌های پردزاش زبان طبیعی در بازیابی اطلاعات به‌طور اعم و در بازیابی اطلاعات موسیقیایی به‌طور اخص وارد شده‌اند تا اجرای این سیستم‌ها را ارتقا بخشند. بررسی‌ها نشان می‌دهد که مدل‌های تجزیه پردازش زبان طبیعی را می‌توان به‌طور موفقیت‌آمیزی برای تجزیه موسیقیایی استفاده کرد. با وجود گستردگی و اهمیت موسیقی در سطح جوامع بشری، حوزه پژوهش درباره بازیابی اطلاعات موسیقیایی نسبتاً جدید است و کمتر از دو دهه از آغاز آن می‌گذرد. بااین‌حال، این حوزه از همان ابتدا تا کنون به‌عنوان یک حوزه پژوهشی، روند رو به جلویی را داشته است. حوزه پژوهش موسیقی با پژوهش در پردازش زبان طبیعی، شباهت دارد. هر دو حیطه بر روی گونه‌های مشابهی از داده‌ها که از ویژگی‌های مشترکی برخوردارند، کار می‌کنند. هر دو حوزه با داده‌هایی کار می‌کنند که انسان‌ها به‌راحتی آن‌ها را درک می‌کنند، اما برای آنکه آن‌ها را به‌طور کامل برای رایانه‌ها قابل‌فهم سازند، با مشکلات فراوانی مواجه‌اند. پژوهشگران موسیقی می‌توانند دستاوردهای خود را در وظایفی مانند جداسازی صداها، یا کشف محدوده‌های اصوات عرضه کنند و در مقابل از پردازش زبان طبیعی در خصوص مواردی همچون روش‌های آماری، رویکردهای خودکار نسبت به معناشناسی، کمک گرفتن از بازیابی اطلاعات و داده‌یابی از طریق فهم زبان طبیعی، استفاده کنند (امیری، ۱۳۹۵).

خلاصه‌سازی خودکار اطلاعات. امروزه با توجه به حجم زیاد اطلاعات و منابع متعدد در اینترنت، وجود ابزارها و روش‌هایی برای خلاصه‌سازی متون ضروری است. خلاصه‌سازی به فرایند تجمیع اطلاعات مفید یک یا چند منبع اطلاعاتی به‌منظور کوتاه کردن متن یا پاسخ‌گویی به درخواست کاربر گفته می‌شود. خلاصه‌سازی خودکار به معنی استفاده از ابزارهای ماشینی و مبتنی بر کامپیوتر برای تولید یک خلاصه مفید و معتبر است و یکی از مسائل مشکل و پر چالش در زمینه پردازش زبان طبیعی به حساب می‌آید، ازآن‌جهت که

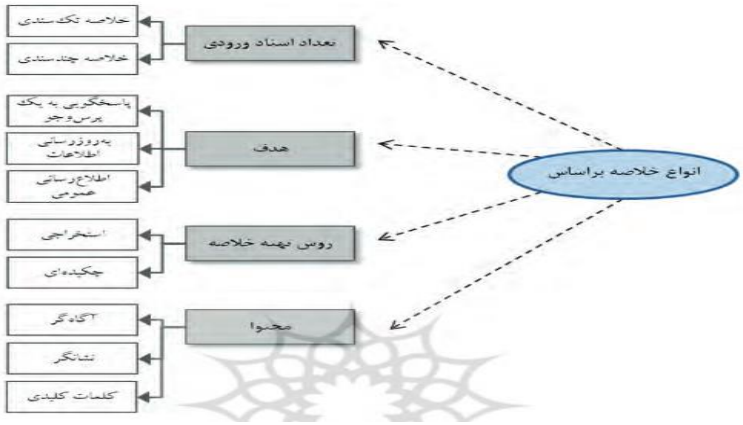
1. Oramas

هنوز کیفیت خلاصه‌های تولیدشده ماشینی به‌اندازه خلاصه‌های انسانی نیست. روش‌های خلاصه‌سازی خودکار به ساختار زبانی متون نگارش شده، در آن زبان وابسته هستند و روش‌های پردازش زبان طبیعی، نقش اصلی را در ایجاد خلاصه‌سازی خودکار ایفا می‌کنند. تا کنون ابزارها و سیستم‌های متعددی برای تولید خلاصه‌ها در زبان انگلیسی ایجاد شده و پژوهش‌های قابل‌توجهی در این زمینه صورت گرفته است؛ که حجم بسیاری از آن‌ها در زمینهٔ خلاصه‌ساز استخراجی است. در زبان فارسی نیز سیستم‌های خلاصه‌ساز استخراجی متنوعی ایجاد شده است که سیستم «فارسی سام» جزء اولین این سیستم‌ها است (ال‌هاشمی[1]، ۲۰۱۰). یکی از مهم‌ترین، چالش‌ها برای یک سیستم خلاصه‌ساز استخراجی، مرحلهٔ پیش‌پردازش است که در آن متن موردنظر برای استخراج جملات خلاصه، عمدتاً بر اساس عملگرهای پردازش زبان طبیعی نظیر ریشه‌یابی، حذف کلمات توقف، برچسب‌زنی نقش کلمات و تعیین کلمات کلیدی پردازش می‌شود. پس‌ازآن، به هر جمله در متن امتیازی تعلق می‌گیرد و درنهایت، جملات با امتیاز بالا انتخاب می‌شوند. ازآنجاکه کلمات کلیدی نقش به سزایی در تعیین امتیاز جملات ایفا می‌کنند، مرحلهٔ پیش‌پردازش اهمیت ویژه‌ای را در خلاصه‌ساز استخراجی دارد. از طرف دیگر، در بین روش‌های خلاصه‌ساز استخراجی، روش‌های مبتنی بر گراف، نتایج اثربخشی ارائه نموده‌اند. لیکن، یکی از چالش‌های اصلی در این روش‌ها، روش پیمایش گراف برای انتخاب جملات مناسب است، به‌گونه‌ای که هر جمله درعین‌حال که لازم است منعکس‌کننده محتوای متن باشد و کلمات کلیدی مناسب را در بر داشته باشد، باید کم‌ترین اشتراک را با جملات قبلی انتخاب‌شده نیز داشته باشد (احمدی، حسینی خواه، محبی، ۱۳۹۶).

انواع خلاصه. خلاصه‌ها را از جهات مختلف می‌توان دسته‌بندی نمود. نوع خلاصهٔ تهیه‌شده بسته به تعداد اسناد یا متون در دست، هدف از خلاصه‌سازی، روش مورداستفاده برای تهیهٔ خلاصه و نوع محتوایی که خلاصه باید در بر داشته باشد، می‌تواند متفاوت باشد.

1. Al-Hashemi

تصویر شماره ۱، دسته‌بندی انواع روش‌های خلاصه‌سازی را بر اساس نوع خلاصه بر مبنای دسته‌بندی «نن‌کووا و مک‌کوون[1]» نشان می‌دهد (نن‌کووا و مک‌کوون، ۲۰۱۲).



تصویر شماره ۱: دسته‌بندی انواع خلاصه (نن‌کووا و مک‌کوون، ۲۰۱۲)

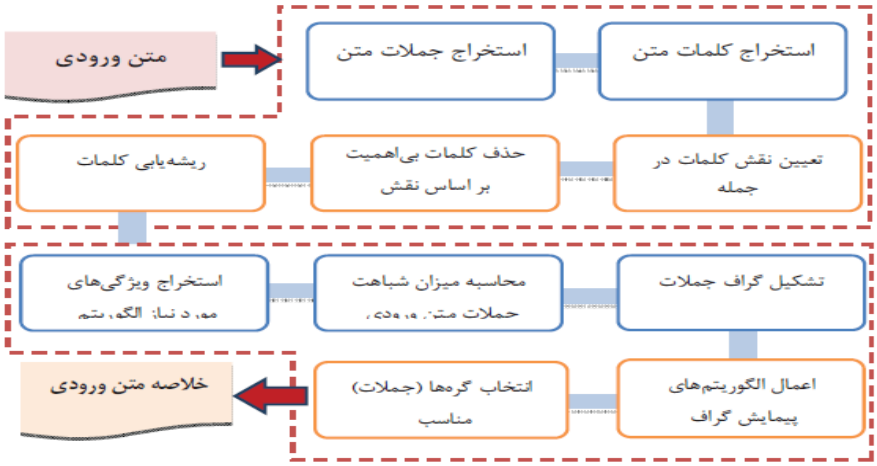
خلاصهٔ تولیدشده می‌تواند حاصل پردازش یک سند یا چندین سند باشد که در اصطلاح به آن خلاصه‌های تک‌سندی یا چندسندی می‌گویند. هدف از تهیهٔ خلاصه می‌تواند در نوع خلاصهٔ تولیدشده مؤثر باشد (کریمی و شمس فرد، ۱۳۸۵). بر اساس تقسیم‌بندی که نن‌کوا و مک‌کوون (۲۰۱۲) در مقالهٔ مروری خود در زمینهٔ روش‌های خلاصه‌سازی متون ارائه کردند، یک خلاصه می‌تواند در راستای پاسخ‌گویی به یک عبارت پرس‌وجو تولید شود. در این حالت مطالبی در خلاصه گنجانده خواهد شد که با عبارت پرس‌وجو مرتبط‌تر است (محامد و راجس‌کاران[2]، ۲۰۰۶). علاوه بر آن خلاصه می‌تواند تنها با هدف به‌روزرسانی یک منبع اطلاعاتی تولید شود، یعنی با فرض موجود بودن مجموعه‌ای از اطلاعات، خلاصه‌ای از اطلاعات به‌روز شده در منبع اضافه شود. درنهایت، یک خلاصه می‌تواند با هدف اطلاع‌رسانی عمومی درباره یک سند یا مجموعه‌ای از اسناد تولید شود، بدون آن‌که اطلاعات خاصی در بدو امر موردنیاز باشد (نن‌کوا و

1. Nenkova & McKeown
2. Mohamed &Rajasekaran

مک‌کوون، ۲۰۱۲). روش تهیهٔ خلاصه می‌تواند بر مبنای استخراج مجموعه‌ای از جملات اصلی باشد که در متون آمده است، یعنی عین جملات متون مرجع در متن خلاصه قید شود که به آن، خلاصهٔ استخراجی می‌گوییم. در نوعی دیگر که عمدتاً چالش‌برانگیزتر است، جملات متون مرجع عیناً آورده نمی‌شود، بلکه جملات جدیدی حاوی خلاصه متن به‌صورت خودکار تولید می‌شود. این حالت، خلاصهٔ چکیده‌ای خوانده می‌شود که عمدتاً روش‌های تولید زبان طبیعی نیز مستقیماً در آن لحاظ خواهد شد. در خلاصهٔ چکیده‌ای متن خلاصه لزوماً تکرار بعضی از جملات متن نیست، بلکه هدف درک متن و تولید جملات جدید و موجز است. به‌عنوان‌مثال، در متن «ساعت حرکت قطار ۱۰ شب بود، اما مریم دیر به ایستگاه رسید و قطار حرکت کرده بود» خلاصهٔ چکیده را می‌توان به شکل زیر بازنویسی کرد: «مریم به قطار ساعت ۱۰ شب نرسید». در این نوع خلاصه، سیستم علاوه بر درک محتوای متن باید قادر به جمله‌سازی بر اساس مطالب اصلی متن باشد، به همین دلیل نسبت به خلاصه‌سازی استخراجی از پیچیدگی بیشتری برخوردار است، اما به خلاصهٔ انسانی (خلاصه‌ای که توسط یک انسان به‌صورت دستی تولید شده است)، شباهت بیشتری دارد (براون[1]، ۲۰۰۲؛ شاکری و همکاران[2]، ۲۰۱۱). درنهایت، یک خلاصه می‌تواند بر اساس محتوایی که دارد، متفاوت باشد. خلاصه نشانگر، معمولاً اطلاعات توصیفی دربارهٔ متن نظیر حجم، شیوه نگارش و موضوع اصلی آن را در بر دارد، درحالی‌که خلاصهٔ آگاه‌گر دربارهٔ محتوای متن اطلاعاتی را ارائه می‌کند. در برخی موارد نیز نیاز است که مجموعه‌ای از کلمات کلیدی مهم یک متن، یا مجموعه‌ای از متون به‌عنوان نماینده محتوای متن استخراج شود که در این صورت خلاصه بر مبنای کلمات کلیدی حاصل می‌شود (دیولا[3]، ۲۰۰۴). به‌عنوان‌مثال چنانچه بخواهیم مراحل اصلی خلاصه‌سازی استخراجی را به‌عنوان یکی از مراحل خلاصه‌سازی خودکار، بررسی نماییم. می‌توان گفت: در هر روش خلاصه‌سازی استخراجی سه مرحلهٔ پیش‌پردازش، پردازش و انتخاب جملات وجود دارد. در مرحله

1. Browne
2. Shakeri, et al.
3. Diola

پیش‌پردازش، روی متن ورودی، عملیات پردازش اولیهٔ متن و نرمال‌سازی آن انجام می‌شود تا برای مراحل بعدی آماده شود. در این مرحله عملگرهای پردازش زبان طبیعی نظیر برچسب‌زنی اجزای گفتار، استخراج کلمات کلیدی و ریشه‌یابی اعمال می‌شود. در مرحلهٔ پردازش با استفاده از مفاهیم آماری، زبان‌شناختی یا ترکیب هر دوی آن‌ها جملات موجود در متن ورودی امتیازدهی می‌گردد. در مرحلهٔ انتخاب جملات، بر اساس معیاری که بر مبنای نیاز کاربر در سیستم تعبیه شده، جملات مناسب انتخاب می‌گردند و درنهایت، جملات مرتب شده و نتیجه به‌عنوان خروجی سیستم خلاصه‌ساز در اختیار کاربر قرار می‌گیرد. مرحله پیش‌پردازش خود شامل فرایندهای استخراج جملات و کلمات، تعیین نقش کلمات در جمله، (برچسب‌زنی کلمات)، حذف کلمات توقف و ریشه‌یابی است (بهروزیان‌نژاد، عطارزاده و حسین‌زاده، ۱۳۹۲) یکی از کارهای اساسی در پردازش زبان طبیعی، برچسب‌زنی اجزای گفتار است. برچسب‌زنی، تعیین مقوله‌های دستوری برای هر نماد در متن است. در برچسب‌زنی از حوزه ساخت واژه و نحو زبان برای تعیین مقوله‌های دستوری استفاده می‌شود. برچسب‌زنی اجزای کلام درواقع، به معنای انتخاب مناسب‌ترین مقولهٔ دستوری به هر کلمه یا ابهام‌زدایی از برچسب کلمه است. بسیاری از واژه‌ها در متون بیش از یک برچسب دستوری دارند، به‌عنوان‌مثال، کلمه تند، در زبان فارسی می‌تواند صفت یا قید باشد. در شکل زیر سه مرحله استخراج خودکار اطلاعات، یعنی پیش‌پردازش، پردازش و انتخاب جملات به تصویر کشیده شده است.



تصویر ۲. معماری بهبود خلاصه‌سازی خودکار متون فارسی با استفاده از روش‌های پردازش زبان طبیعی و گراف شباهت (حسینی‌خواه، احمدی و محبی، ۱۳۹۵)

مراحل نشان داده‌شده در شکل را می‌توان به دو بخش کلی تقسیم نمود:

مرحلهٔ پیش‌پردازش: شامل مراحل استخراج جملات و کلمات، تعیین نقش کلمات در جمله، حذف کلمات توقف و ریشه‌یابی کلمات

مرحلهٔ پردازش: شامل مراحل استخراج ویژگی‌های کلمات هر متن، محاسبهٔ میزان شباهت همهٔ جملات یک متن به هم، ساخت گراف برای هر متن ورودی، اعمال.

الگوریتم‌های پیمایش گراف و درنهایت، انتخاب گره‌های (جملات) مهم و نمایش آن در خروجی (ریجسبرگن[1]، ۱۹۷۹).

یکی از معیارهای ارزیابی که در اغلب کاربردهای پردازش زبان طبیعی استفاده می‌شود، معیار دقت و بازخوانی است. دقت برابر است با نسبت تعداد جملات درستی که توسط سیستم خلاصه‌ساز انتخاب‌شده به کل جملاتی که سیستم برای خلاصه ارائه کرده است. بازخوانی برابر است با نسبت تعداد جملات درستی که توسط سیستم خلاصه‌ساز انتخاب‌شده به کل جملاتی که توسط انسان برای خلاصه ارائه شده است. این دو معیار برای سنجش کارکرد

1. Rijsbergen

سیستم خلاصه‌ساز بسیار مناسب هستند. علاوه بر کارایی سیستم خلاصه‌ساز، متن ورودی و موضوع متن نیز بر کیفیت خلاصه تأثیرگذار است، چراکه منابع خبری مختلف در رعایت استانداردهای نگارشی نیز با هم متفاوت هستند. به همین ترتیب انواع موضوعات نیز، در کیفیت خلاصه تأثیرگذار است. به‌عنوان‌مثال، یک خبر ورزشی اغلب دارای جملات کوتاه و اسامی خاص بیشتری، نسبت به یک خبر اجتماعی است و همین مسائل، بر کیفیت خلاصه سیستمی نیز تأثیرگذار است (مانینگ و اسچاتز[1]، ۱۹۹۹).

نمایه‌سازی خودکار اطلاعات. همچنین، یکی دیگر از کاربردهای پردازش زبان طبیعی ساخت نمایه‌سازی خودکار است. یکی از مشکل‌ترین و مهم‌ترین وظایف سیستم‌های بازیابی متن امکان جستجوی کارا روی اطلاعات متنی است. دراین‌بین، گسترش روزافزون منابع اطلاعات علمی، باعث گرایش متخصصان اطلاعات به فشرده‌گویی و استفاده از راهکارهای آسان‌سازی جست‌وجوی اطلاعات شده است. دراین‌بین نمایه‌سازی، یکی از باصرفه‌ترین راه‌های میان‌بر جهت رسیدن به اطلاعات است (تشکری و میبدی، ۱۳۸۲). در جستجو و بازیابی اطلاعات، یک نمایه مناسب در حقیقت نقش کلیدی دارد. ساخت نمایه برای متن‌ها، بازیابی آن‌ها را برای محققان و خوانندگان آسان می‌سازد. عمل ساخت نمایه در حال حاضر به دست نیروی انسانی متخصص و ماهر انجام می‌شود که این روش دشوار، پرهزینه و زمان‌بر است؛ اما در سیستم‌های بازیابی متن، چنین نمایه‌ای می‌تواند کاملاً به‌صورت خودکار تولید شود. به‌عبارت‌دیگر یک سیستم بازیابی متن، می‌تواند مجموعه اقداماتی را در جهت ساخت یک نمایه مناسب، مجهز و کارآمد روی واژه‌های متن انجام دهد. پس از ساخت نمایه، سیستم می‌تواند با استفاده از آن، در جواب به پرس‌وجوی کاربر، متونی را که مربوط به واژه‌های مورد درخواست وی هستند را یافته و ارائه کند (خالویی، ۱۳۸۵). در بین روش‌های بی‌شمار نمایه‌سازی که هر یک دارای نقاط ضعف و قوت خاص خود هستند، نمایه‌سازی خودکار یکی از روش‌هایی است که علاوه بر جست‌وجو پذیر نمودن اطلاعات موجود، باعث افزایش توان آدمی در برابر پدیده انفجار اطلاعات و افزایش بی‌وقفه داده‌ها،

1. Manning & Schutze

به‌خصوص در قالب الکترونیکی شده است. با پیشرفت و تکامل نرم‌افزارهای نمایه‌سازی خودکار در افزایش سرعت فرایند نمایه‌سازی، تلفیق راهکارهای نمایه‌سازی خودکار و دستی و کاهش زمان و هزینه‌ها، اقبال و گرایش مدیران پایگاه‌های اطلاعاتی جهت استفاده از این فناوری‌ها بیش از بیش افزایش یافته است (گیلوری، ۱۳۷۹). تعاریف چندی برای نمایه‌سازی خودکار ارائه شده است: فرایند استخراج مجموعه‌ای از مدخل‌های نمایه‌ای که بیانگر موضوع متن هستند توسط رایانه از متن ماشین‌خوان را نمایه‌سازی خودکار می‌نامند. اگر نمایه‌سازی را فرایندی دومرحله‌ای فرض کنیم، شامل: انتخاب و استخدام واژگان که بازنمون مناسبی از اطلاعات باشند و همچنین جست‌وجو پذیر ساختن واژگان انتخاب‌شده، هر نوع نمایه‌سازی را که هر دوی این مراحل و یا حداقل یکی از مراحل فوق را به‌صورت خودکار و ماشینی ارائه دهد، نمایه‌سازی خودکار می‌نامند. طبق تعریفی دیگر نمایه‌سازی خودکار عبارت است از نوعی از نمایه‌سازی که از الگوریتم‌های اجرایی در محیط‌های الکترونیکی پیروی می‌کند. این الگوریتم‌ها در پایگاه‌های داده با اطلاعات کتابشناختی و حتی متن کامل اجرا می‌شوند. همچنین پایگاه‌های غیرمتنی مانند پایگاه‌های تصویری و یا موسیقی (صوتی) نیز قابل نمایه‌سازی با این الگوریتم‌ها می‌باشند. نمایه‌ها دارای کارکردهای زیر می‌باشند.

- محتوای اطلاعاتی مدارک را فشرده می‌سازند.
- به‌عنوان واسطه‌ای برای تطبیق و یکسان‌سازی زبان مدرک و زبان کاوش به کار می‌روند.
- به‌عنوان ابزاری کارآمد بر شیوه تدوین راهبردهای کاوش در جست‌وجوهای اطلاعاتی نظارت دارند. (همیدی، کانان و ایونس[1]، ۱۹۹۷)

نمایه‌سازی پایگاه‌های اطلاعاتی و نمایه‌های ماشینی به روش‌های زیر ممکن است انجام گیرد.

- پایگاه‌های اطلاعاتی

1. Hmeidi, I., Kanaan, G., and Evens

- بهره‌گیری از رایانه برای کنترل کیفیت نمایه‌های تولیدی مثل بررسی این مسئله که آیا همهٔ اصطلاحات نمایه در اصطلاح‌نامه وجود دارند یا خیر؟
- بهره‌گیری فکری از رایانه مثل استفاده از رایانه، برای مثال استفاده از رایانه برای وزن‌دهی و انتخاب اصطلاحات نمایه‌ای
- نمایه‌سازی کامل خودکار به کمک رایانه
- در تاستفاده از رایانه برای انجام امور دفتری نمایه‌سازی مثل ورود اطلاعات در
- مامی این موارد از تکنیک‌های پردازش زبان طبیعی استفاده می‌شود.

لیدی[1] که در زمینهٔ سطوح مختلف نمایه‌سازی رایانه‌ای به تحقیق پرداخته است، بیان می‌دارد اصول پردازش زبان طبیعی که آن‌ها را زبان‌های سطح پایین می‌گوید (اصوات، کلمات و عبارات اسمی) بیشتر با نمایه‌سازی خودکار عجین‌اند، درحالی‌که زبان‌های سطح بالا که شامل معانی، زبان‌شناسی، واقع‌گرایی و تعامل است، بیشتر با روح نمایه‌سازی انسانی سازگار است. (دولانی و فرهادپور، ۱۳۸۸) در همین راستا امروزه شرکت‌ها و ناشران رسانه‌های الکترونیکی، دارندگان صفحات وب و سرویس‌دهندگان شبکه‌های اینترنتی در پی نمایه‌سازی مطلوب صفحات فرامتن خود هستند. آن‌ها نرم‌افزارهایی را در جهت نمایه‌سازی خودکار ارائه کرده‌اند که این نرم‌افزارها بر پایه تکنیک‌های پردازش زبان طبیعی و به‌صورت خودکار به ساخت نمایه‌سازی به‌صورت خودکار اقدام می‌نمایند. چند مورد از مشهورترین این نرم‌افزارها عبارت‌اند از: نرم‌افزار پرپ/اچ‌تی‌ام‌ال[2]، نرم‌افزار اچ‌تی‌ام‌ایندکسر[3]، نرم‌افزار روبوهلپ اچ‌تی‌ام‌ال اِدیشن[4]، نرم‌افزار ماکرِکس[5]، نرم‌افزار سیندِکس[6]، نرم‌افزار ریترایورِ[7] و نرم‌افزار اِسکای ایندِکس[8] (دولانی و فرهادپور، ۱۳۸۸)

1. Liddy
2. Prep/HTML
3. HTML indexer
4. RoboHelp HTML Edition
5. Macrex
6. Cindex
7. Retriever
8. Sky index

تفاوت نمایه‌سازی دستی و نمایه‌سازی خودکار. علاوه بر دشواری، کندی و پرهزینه بودن نمایه‌سازی در شیوه دستی، این شیوه مشکلات دیگری نیز در بر دارد. در شیوه دستی هر فرد نمایه‌ساز که به‌خوبی هم آموزش دیده است نسبت به فرد دیگر نمایه‌های متفاوتی، به یک متن می‌دهد. حتی دیده شده است که یک نمایه‌ساز نیز در زمان‌های مختلف نمایه‌های متفاوتی به یک متن می‌دهد. اضافه بر این هرچند افراد نمایه‌ساز درک بهتری از متن دارند، ولی ممکن است هنگامی‌که با حجم زیادی از متون و واژه‌های نمایه سروکار دارند، اشتباهات زیادتری نسبت به شیوه‌های خودکار نمایه‌سازی داشته باشند. بااین‌وجود روند نمایه‌سازی خودکار بهتر است تنها روی مجموعه‌های متنی بزرگ به کار رود. در مورد مجموعه‌های کوچک، این امکان وجود دارد که روش‌های دستی سریع‌تر بوده و نتیجه پرکاربردتری ارائه کنند. اصولاً هر متنی دارای دو بخش است: بخش اول شامل اطلاعاتی است که نسبت به محتوای متن خارجی محسوب می‌شوند، مثل نام نویسنده، تاریخ و محل انتشار و نام منتشرکننده؛ بخش دوم شامل محتویات اصلی متن است (کسلج[1]، ۲۰۰۳). در کتابخانه‌ها بخش اول با نام دسته‌بندی توصیفی و بخش دوم با نام دسته‌بندی موضوعی شناخته می‌شود. نمایه‌سازی به عملیات شناخت موضوع و محتوای متن گفته می‌شود و هنگامی‌که این عملیات به کمک وسایل مدرن کامپیوتری و نیز به کمک تکنیک‌های پردازش زبان طبیعی صورت پذیرد، نمایه‌سازی خودکار نامیده می‌شود. حساس‌ترین و مشکل‌ترین مرحله‌ای که در روند نمایه‌سازی خودکار باید طی شود، انتخاب واژه‌هایی است که برای ساخت نمایه به کار می‌روند. در عمل، نمایه‌سازی روی تمام واژه‌های متن دارای سربار بسیار زیاد است، ضمن اینکه نمایه‌سازی روی تمام واژه‌ها یک کار غیرضروری است. کافی است تنها واژه‌هایی در نمایه به کار روند که نشان‌دهنده محتوای متن مربوطه هستند، درواقع واژه‌ها یا مفاهیمی که موردعلاقه کاربر بوده، توسط وی جستجو می‌شوند. اولین و واضح‌ترین مکانی که شناساننده‌های خوب محتوا ممکن است یافت شود، خود متن مستندات است (کُپ[2]، ۱۹۹۲).

1. Keselj
2. Cope

مراحل نمایه‌سازی خودکار. الکساندرکایزر[1] (در مقاله خالویی) نمایه‌سازی به کمک رایانه را راهی برای پیوند دادن نمایه‌سازی دستی و ماشینی معرفی می‌کند و این فرایند را به سه بخش: ۱ - تحلیل متن ۲- استفاده از روش‌های نمایه‌سازی ماشینی و ۳- کنترل به‌وسیله یک نمایه‌ساز متخصص تقسیم می‌نماید (خالویی، ۱۳۸۵).

شرایط نمایه‌سازی خودکار. به‌طورکلی قبل از شروع ساخت نمایه باید به بعضی موارد مهم توجه داشت که به مهم‌ترین این موارد در ذیل اشاره می‌شود. در نمایه‌سازی دستی وسایل مختلفی وجود دارد تا به نمایه‌ساز در کنترل عملیات نمایه‌سازی کمک نماید. واضح است که این‌گونه وسایل که به شکل متون زبان طبیعی هستند به‌سادگی نمی‌توانند در یک سیستم نمایه‌سازی خودکار مورداستفاده قرار گیرند. نمایه‌سازی می‌تواند به دو روش کنترل‌شده و کنترل نشده انجام شود. در روش نمایه‌سازی کنترل نشده، همه واژه‌های متن در ساخت نمایه به کار می‌روند. این کار معمولاً علاوه بر ایجاد سربار زیاد، می‌تواند باعث ایجاد کج‌فهمی و خطا در روند جستجوی کاربر شود؛ بنابراین روش نمایه‌سازی کنترل‌شده که در آن واژه‌های نمایه کاملاً محدود هستند، غالباً موردحمایت قرار می‌گیرد. در روش کنترل‌شده صحت املای واژه‌ها بررسی می‌شود و نیز برای واژه‌های مترادف یک واژه واحد انتخاب می‌شود. مثلاً واژه «نوسان» نماینده واژه‌های «ارتعاش»، «تموج»، «نوسان»، «تپش»، «چرخش» و غیره می‌گردد. لکن استفاده از این نوع نمایه باعث می‌گردد تا اشخاص آموزش‌دیده‌ای برای فرموله کردن جملات پرس‌وجو موردنیاز باشد. از این روش در ساخت نمایه به‌صورت دستی استفاده می‌شود. شواهد نشان می‌دهد که در صورت استفاده از روش نمایه‌سازی کنترل نشده نتایج نزدیک به روش کنترل‌شده خواهد بود. در نمایه‌سازی می‌توان از واژه‌های ساده و یا مرکب استفاده کرد. در روش پس‌همارایی[2] از واژه‌های ساده برای ساخت نمایه استفاده می‌شود و در موقع جستجو کاربر می‌تواند با ترکیب واژه‌های ساده و بر اساس نمایه موارد متنی موردنظرش را بیابد. این روش بیشتر در سیستم‌هایی که

1. Kaiser
2. postcoordination

ساخت نمایه به‌صورت خودکار انجام می‌شود، به کار می‌رود. در روش پیش‌همارایی[1] واژه‌های ترکیبی هم در ساخت نمایه به کار می‌روند. از این روش بیشتر در سیستم‌هایی که نمایه به‌صورت دستی ساخته می‌شود، استفاده می‌شود. در سیستم‌های نمایه‌سازی خودکار غالباً از واژه‌های مجزا استفاده می‌شود، زیرا انتساب خودکـار این‌گونه از واژه‌ها به متون بهتر انجام می‌شود. در ساخت نمایه‌سازی خودکار ازجمله نکاتی که می‌بایست مدنظر قرار بگیرد، حذف واژه‌هایی مانند ضمایر، قیود، حروف اضافه و ربط هستند که کاربر مایل به جستجوی آن‌ها نیست؛ بنابراین جهت نمایه‌سازی خودکار ابتدا سیستم باید این واژه‌ها را از بین واژه‌های استخراج‌شده از متون حذف کند. (تشکری و میبدی، ۱۳۸۲؛ گیلوری، ۱۳۷۹) همچنین، برای ارزیابی کیفیت بازیابی در سیستم‌های بازیابی متنی ازجمله سیستم‌های نمایه‌سازی خودکار، پارامترهای زیادی تعریف‌شده‌اند. مهم‌ترین این پارامترها عبارت‌اند از: بازخوانی و دقت. بازخوانی بیانگر قابلیت سیستم برای ارائه موارد مربوط به درخواست کاربر است. درواقع ایـن پـارامتر را می‌توان به‌صورت زیر تعریف کرد:

$$\text{"بازخوانی"} = \frac{\text{تعداد موارد بازیابی شده و مربوط}}{\text{تعداد کل موارد مربوط در متن}}$$

مقدار این پارامتر همواره بین صفر و یک است و نسبت مستقیمی با میزان دربرگیری واژه‌ها در نمایه دارد؛ یعنی هر چه نمایه واژه‌های بیشتری را در خود داشته باشد، تعداد موارد بازیابی شده مربوط («بازخوانی») بیشتر می‌شود. از طرف دیگر میزان این پارامتر به مجموعه متون، مجموعه کاربران و نحوه فرموله کردن پرس‌وجو توسط کاربران بستگی دارد.

دقت: این پارامتر به بیان ساده بیانگر قابلیت سیستم برای ارائه فقط موارد مربوطه به درخواست کاربر (یا به‌عبارت‌دیگر رد موارد نامربوط) است. درواقع این پارامتر را نیز می‌توان به‌صورت زیر تعریف کرد:

1. precoordination

$$\text{"دقت"} = \frac{\text{تعداد موارد بازیابی شده و مربوط}}{\text{تعداد کل موارد بازیابی شده}}$$

مقدار این پارامتر نیز بین صفر و یک است و نسبت مستقیمی با میزان تعیین‌کنندگی واژه‌ها در نمایه دارد؛ یعنی هرچقدر واژه‌های به‌کاررفته در نمایه دقیق‌تر و تعیین‌کننده‌تر باشند، تعداد موارد نامربوط کم‌تری بازیابی می‌شوند. از طرف دیگر میزان این پارامتر نیز به مجموعه متون، مجموعه کاربران و نحوه فرموله کردن پرس‌وجو توسط کاربر هم‌بستگی دارد. در حالتی که تعداد کل موارد بازیابی شده صفر است، نمی‌توان مقدار «دقت» را با استفاده از فرمول بالا محاسبه نمود. منطقاً در این حالت مقدار پارامتر «دقت» یک در نظر گرفته می‌شود. مقدار دو پارامتر «بازخوانی» و «دقت» غالباً نسبت معکوس با هم دارند و بهبود یکی موجب افت دیگری می‌شود (مهراد و ناصری، ۱۳۸۷).

سیستم‌های پرسش و پاسخ (کوِزشِن اَسک سیستم[1])، شکل پیچیده‌تری از بازیابی اطلاعات است. سیستم‌های هوشمند پرسش و پاسخ خودکار نتایج پرسش‌های کاربران را در قالب پاسخی مختصر و صریح ارائه می‌دهند. در چند دهه اخیر پیشرفت‌های زیادی در توسعه این سیستم‌ها به وجود آمده و چندین سیستم قدرتمند توسعه داده شده است. بااین‌حال متأسفانه اقدامات کمی در این حیطه برای زبان فارسی صورت گرفته است. در این سیستم‌ها سؤال به زبان طبیعی و بدون هیچ محدودیت معنایی به‌عنوان ورودی به سیستم داده می‌شود. وظیفه سیستم یافتن جوابی دقیق، کوتاه و کامل برای سؤال داده شده در کوتاه‌ترین زمان ممکن است.

به این منظور سیستم‌های کوِزشِن اَسک، تکنیک‌های بازیابی اطلاعات، استخراج اطلاعات و پردازش زبان طبیعی را با هم به کار می‌گیرند. نمونه چنین سیستمی در کتابداری و اطلاع‌رسانی، خدمات مرجع از کتابدار بپرس هست. در این نوع خدمات کاربر سؤالاتی را به زبان طبیعی از کتابدار می‌پرسد و کتابدار مرجع، پاسخ سؤالات را به‌صورت دقیق و با در نظر گرفتن اولویت‌های زمانی برای کاربر ارسال می‌کند که همان‌گونه که ذکر شد در

1. Question & Ask Systems

این سیستم‌ها از تکنیک‌های پردازش زبان طبیعی و بازیابی اطلاعات با هم استفاده می‌گردد. هدف پرسش و پاسخ این است که به جست‌وجوها پاسخ‌های رضایت‌بخشی ارائه دهد. نیازهای اطلاعاتی باید به‌خوبی تعریف شود. در اینجا پردازش زبان طبیعی می‌کوشد تا نوع پاسخ را با شفاف‌سازی سؤال، تحلیل مجموعه محدودیت‌ها و با استفاده از فنون استخراج اطلاعات تعیین کند. این نظام‌ها در نظر است جانشینان احتمالی نظام‌های بازیابی اطلاعات کنونی شوند. علاوه بر خدمات مرجع از کتابدار بپرس، نظام زبان طبیعی اسمارت نیز نمونه‌ای از این سیستم‌ها است (سرلک، خلجی و گردان، ۱۳۹۵).

رده‌بندی خودکار متون. درنهایت باید به فنون خودکار رده‌بندی متن اشاره کنیم که به‌طور خودکار مجموعه‌ای از مدارک را به مقوله‌هایی در چارچوب رده‌بندی‌های از پیش تعریف‌شده نسبت می‌دهند. در این رابطه توصیف صحیح ویژگی‌های مدرک، قویاً کیفیت گروه‌بندی و یا مقوله‌بندی را به‌وسیله این فنون متأثر ساخته است (بهروزیان نژاد، عطار زاده و حسین زاده، ۱۳۹۲).

نتیجه‌گیری

به‌طورکلی و بر اساس مطالب پیش‌گفته می‌توان گفت که مزایای خوبی در استفاده از فنون پردازش زبان طبیعی در کاربردهای کتابداری و اطلاع‌رسانی و ازجمله بازیابی اطلاعات مشاهده می‌کنیم و می‌توان گفت موفقیت یک نظام بازیابی اطلاعات برای کاربر وابسته به این مطلب است که تا چه حد اصول پردازش زبان طبیعی در آن مورداستفاده قرار می‌گیرد. ولی باوجوداین، باید توجه داشت که این مزایا در پرتو هزینه‌های زیاد محاسباتی امکان‌پذیر است و گاهی در برخی از موارد مشاهده‌شده که فنون غیر از پردازش زبان طبیعی پیشرفت‌های زیادتری به بار می‌آورد (مهراد و ناصری، ۱۳۸۷). درعین‌حال پردازش زبان طبیعی، همچنان دارای قابلیت‌های خوب و مفیدی در حوزه‌های مختلف و ازجمله در رشته علوم کتابداری و اطلاع‌رسانی دارد که می‌بایست با برشمردن مزایا و هزینه‌ها نسبت به ادغام پردازش زبان طبیعی در حوزه‌های موضوعی مختلف اقدام نمود.

منابع

احمدی، عباس؛ حسینی‌خواه، طیبه و محبی، آزاده. (۱۳۹۶). بهبود خلاصه‌سازی خودکار متون فارسی با استفاده از روش‌های پردازش زبان طبیعی و گراف شباهت. *فصلنامه پردازش و مدیریت اطلاعات*، ۳۳(۲)، ۹۱۴-۸۸۵.

امیری، ناهید. (۱۳۹۵). پردازش زبان طبیعی و بازیابی اطلاعات موسیقیایی. *ششمین همایش پژوهش‌های نوین در علوم و فناوری*

بهروزیان‌نژاد، محمد؛ عطارزاده، ایمان و حسین‌زاده، مهدی. (۱۳۹۲). مقایسه روش‌های دسته‌بندی خودکار متون. *اولین همایش ملی رویکردهای نوین در مهندسی کامپیوتر و بازیابی اطلاعات*.

تشکری، مسعود و میبدی، محمدرضا. (۱۳۸۲). ساخت یک نمایه‌ساز خودکار برای متون فارسی. *یازدهمین کنفراس مهندسی برق*. بازیابی از: http://ce.aut.ac.ir/~meybodi/paper/Tashakori-meybodi-Automatic%20indexer-11-ICEE-Shiraz-1382.pdf

خالویی، مرضیه. (۱۳۸۵). نمایه‌سازی ماشینی. *نما*. ۶(۳).

دولانی، عباس و فرهادپور، محمدرضا. (۱۳۸۸). مروری بر نمایه‌سازی خودکار و نرم‌افزارهای رایج تولید آن. *فصلنامه کتاب*، ۲۰(۳)، ۳۱۰-۲۹۱.

سرلک، ولی؛ خلجی، مجید و گردان، محمد. (۱۳۹۵). بررسی سیستم‌های هوشمند پرسش و پاسخ خودکار زبان فارسی با استفاده از اطلاعات وب جهانی دانشنامه رشد و ویکی‌پدیا. مقاله کنفرانس. *دومین کنفرانس بین‌المللی یافته‌های نوین پژوهشی در علوم، مهندسی و فناوری*.

سزاوار، امیر؛ فرسی، حسن و محمدزاده، سجاد. (۱۳۹۵). بازیابی تصویر با استفاده از یادگیری عمیق. *چهارمین کنفرانس بین‌المللی پژوهش‌های کاربردی در مهندسی کامپیوتر و پردازش سیگنال*. قابل‌دسترس در https://www.civilica.com/Paper-CEPS ۰٤-۰٤CEPS_۰۷۳

کریمی، زهره و شمس‌فرد، مهرنوش. (۱۳۸۵). سیستم خلاصه‌ساز خودکار متون فارسی. *دوازدهمین کنفرانس بین‌المللی انجمن کامپیوتر ایران*. تهران

گیلوری، عباس. (۱۳۷۹). نمایه‌سازی خودکار: گذشته، حال، آینده. *فصلنامه پیام کتابخانه*، ۱۰(۴)، ۲۵-۱۵.

علیزاده، حمید. (۱۳۸۳). مشکلات دسترسی به اطلاعات در جهان شبکه‌ها. *فصلنامه کتاب*. ۱۵(۲)، ۱۲۱-۱۱۵.

علیزاده، حمید؛ فتاحی، رحمت‌الله و داورپناه، محمدرضا. (۱۳۸۸). بررسی کارآمدی روش‌های موجود در بازیابی اطلاعات بین‌زبانی فارسی – انگلیسی با استفاده از واژه‌نامه دوزبانه ماشین‌خوان. *فصلنامه پردازش و مدیریت اطلاعات*، ۲۵(۱)، ۷۰-۵۳.

مهراد، جعفر و ناصری، مریم. (۱۳۸۷). *پردازش زبان طبیعی و بازیابی اطلاعات*. تهران: چاپار- شیراز: مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری.

پردازش زبان طبیعی چیست. بازیابی از: http://jamejamonline.ir/Media/pdfs/1390/05/16/100850852371.pdf

References

Adriani, M. (2000). Using statistical term similarity for sense disambiguation in cross-language information retrieval. *Information retrieval*, *2*(1), 71-82.

Al-Hashemi, R. (2010). Text Summarization Extraction System (TSES) Using Extracted Keywords" *Int. Arab J. e-Technol*. 1 (4), 164–168.

Bod, R. (2001). Memory-Based Models of Music Analysis: Evidence against the Gestalt Principles in Music. *Proceedings International Computer Music Conference*, Havana, Cuba.

Bod, R. (2002). A unified model of structural organization in language and music. *JAIR*, 7(1), 289–308.

Bod, R. (2012). *Probabilistic Grammars for Music*. ILLC: University of Amsterdam

Brants, T. (2003). Natural Language Processing in Information Retrieval. *CLIN*, *111*.

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544-551.

Browne, G. (2002). Automatic indexing and abstracting. *Journal of the American Society for Information Society*. [on-line]. Available: http://www. autoindexing\automatic indexing and abstracting.htm

Chen, H., Schatz, B., Ng, T., Martinez, J., Kirchhoff, A., & Lin, C. (1996). A parallel computing approach to creating engineering concept spaces for semantic retrieval: The Illinois Digital Library Initiative Project. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, *18*(8), 771-782.

Cope, D. (1992). Computer modeling of musical intelligence in EMI. *Computer Music Journal*, *16*(2), 69-83.

Croft, W. B. (1995). NSF center for intelligent information retrieval. *Communications of the ACM*, *38*(4), 42-43.

Diola, A. M., Lopez, J. T. T. O., Torralba, P. F., So, S., & Borra, A. (2004). Automatic Text Summarization. In *Proceedings of the 2 nd National Natural Language Processing Research Symposium*.

Hjørland, B. (2002). Epistemology and the socio-cognitive perspective in information science. *Journal of the American Society for Information science and Technology*, *53*(4), 257-270.

Hmeidi, I., Kanaan, G., & Evens, M. (1997). Design and implementation of automatic indexing for information retrieval with Arabic documents. *Journal of the American Society for Information Science*, *48*(10), 867-881.

Kaiser, A. (1993). *Computer Unterstue Ztes Indexieren in Intelligeten Information Retrieval systemen*. Doktorat Dissertation. Fachbereich Informationswirt schaft, Wirtschafts universitaet Wien.

Kešelj, V., Peng, F., Cercone, N., & Thomas, C. (2003, August). N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING* (Vol. 3, pp. 255-264). sn.

Fred, L., & Ray, J. (1983). *A generative theory of tonal music.* The MIT Press.

Lewis, D. D., & Jones, K. S. (1996). Natural language processing for information retrieval. *Communications of the ACM*, *39*(1), 92-101.

Liddy, E. D. (2003). Natural language processing. lnEncylopedia of library and information science 2nd. New York: Marcel Dekker.

Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Mohamed, A. A., & Rajasekaran, S. (2006, August). Improving query-based summarization using document graphs. In *2006 IEEE international symposium on signal processing and information technology* (pp. 408-410). IEEE. Available at:https://www.researchgate.net/publication/232638359_Improving_Query-Based_Summarization_Using_Document_Graphs.

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544-551. Available at: http://jamia.oxfordjournals.org/content/18/5/544

Nenkova, A., & McKeown, K. (2012). A survey of text summarization techniques. In *Mining text data* (pp. 43-76). Springer, Boston, MA.

Oramas, Sergio (2014). *Harvesting and Structuring Social Data in Music Information Retrieval*. Barcelona, Spain: Universitat Pompeu Fabra.

Van Rijsbergen, C. (1979). Information retrieval: theory and practice. In *Proceedings of the Joint IBM/University of Newcastle upon Tyne Seminar on Data Base Systems* (pp. 1-14). http://www.dcs.gla.ac.uk/Keith

Salton, G. & Mc Gill, M. J (1983*). Introduction to Modern Information Retrieval*, Mc Graw Hill, New York.

Sanderson, M. (2000). Retrieving with good sense. *Information retrieval*, *2*(1), 49-69.

Schedl, M. (2013). *On the Use of the Web and Social Media in Multimodal Music Information Retrieval*. Postdoctoral Thesis (Habilitation).

Shakeri, H., Gholamrezazadeh, S., Salehi, M. A., & Ghadamyari, F. (2012). A new graph-based algorithm for Persian text summarization. In *Computer science and convergence* (pp. 21-30). Springer, Dordrecht.

Taghva, K., Beckley, R., & Sadeh, M. (2005, April). A stemming algorithm for the Farsi language. In *International Conference on Information Technology: Coding and Computing (ITCC'05)-Volume II* (Vol. 1, pp. 158-162). IEEE.

Temperley, D. (2007). *Music and Probability*. The MIT Press

Wang, J., & Oard, D. W. (2005, September). Clef-2005 cl-sr at maryland: Document and query expansion using side collections and thesauri. In *Workshop of the Cross-Language Evaluation Forum for European Languages* (pp. 800-809). Springer, Berlin, Heidelberg. http://www.informatik.uni-trier.de/~ley/db/conf/clef/clef2005.html#WangO05