نشریه الکترونیکی سازمان کتابخانه‌ها، موزه‌ها و مرکز اسناد آستان قدس رضوی
شاپا الکترونیکی: ۴۵۱X-۲۰۰۸

اِن اِل پی آی آر: چهارچوبی نظری برای به کارگیری پردازش زبان طبیعی در بازیابی اطلاعات[1]

مترجم: مهدی زینالی تازه کندی

چکیده

نقش بازیابی اطلاعات در پشتیبانی از تصمیم‌گیری و مدیریت دانش به موضوع قابل‌توجهی تبدیل شده است. برخلاف مشکلات مختلف بازیابی اطلاعات سنتی مبتنی بر کلیدواژه، بسیاری از پژوهشگران در مورد پتانسیل فناوری‌های پردازش زبان طبیعی پژوهش کرده‌اند. با وجود کاربرد گسترده پردازش زبان طبیعی در بازیابی اطلاعات و انتظارات زیاد مبنی بر این که پردازش زبان طبیعی می‌تواند مشکلات بازیابی اطلاعات سنتی را برطرف کند؛ توسعه یک جزء پردازش زبان طبیعی برای یک نظام بازیابی اطلاعات هنوز هم فاقد چهارچوب منسجم پشتیبانی‌کننده است. در این مقاله یک چهارچوب نظری با عنوان «اِن اِل پی آی آر» ارائه شده است که هدف آن ادغام پردازش زبان طبیعی در بازیابی اطلاعات و تعمیم کاربرد گسترده پردازش زبان طبیعی در آن است. برخی از فنون پردازش زبان طبیعی موجود برای تأیید چهارچوب ارائه شده است که نه تنها می‌تواند برای پژوهش‌های فعلی به کارگرفته شود، بلکه پیش‌بینی‌شده است که از پژوهش‌ها و توسعه آینده بازیابی اطلاعات که شامل پردازش زبان طبیعی است، پشتیبانی کند.

کلیدواژه: فنون پردازش زبان طبیعی، بازیابی اطلاعات، تحلیل معنا، تکواژ شناسی، موتور کاوش.

سازمان کتابخانه‌ها، موزه‌ها و مرکز اسناد آستان قدس رضوی
نشریه الکترونیکی شمسه، دوره ۱۲، شماره ۴۶- ۴۷، بهار و تابستان ۱۳۹۹، صص. ۷۴-۹۴.

1. Zhou, L., & Zhang, D. (2003). NLPIR: A theoretical framework for applying natural language processing to information retrieval. *Journal of the American Society for Information Science and Technology*, 54(2), 115-123.

مقدمه

از آنجا که ارزش اطلاعات الکترونیکی به طور چشمگیری افزایش یافته و شبکه‌های رایانه‌ای رایج‌تر شده است، اغلب مردم از موتورهای کاوش برای یافتن اطلاعات مورد علاقه خود در اینترنت یا اینترانت استفاده می‌کنند. با وجود این، افراد به هنگام بازیابی اطلاعات با مشکلات مختلفی روبه‌رو می‌شوند که ناشی از حجم زیاد اطلاعات است و به طور فزاینده‌ای از توانایی پردازش انسان فراتر رفته است. حتی موتورهای کاوش به اصطلاح هوشمند به حد کافی کارا نیستند. باگذشت زمان، نیاز به بازیابی اطلاعات به موقع و دقیق برای پشتیبانی از تصمیم‌گیری و مدیریت دانش به طور قابل‌توجهی بیشتر شده است. در این مقاله، ما به بازیابی اطلاعات مبتنی بر متن می‌پردازیم. زبان طبیعی سرشار از ابهام و غنا به هنگام اظهار در تمام سطوح سازه‌های زبانی است. این امر عملکرد نظام‌های بازیابی اطلاعات، یعنی روش‌های درک پرسش، تجزیه و تحلیل مدارک، مطابقت مدارک با پرسش و تعیین شباهت‌های آن‌ها را به چالش می‌کشد. بنابراین، بسیاری از پژوهشگران در حال بررسی پتانسیل‌های فنون پردازش زبان طبیعی هستند. با وجود کاربردهای گسترده پردازش زبان طبیعی در بازیابی اطلاعات و موارد بسیار زیاد انتظارات از بهبود عملکرد پردازش زبان طبیعی، هنوز هم هیچ چهارچوب منسجمی برای پشتیبانی از پژوهش و توسعه اجزای پردازش زبان طبیعی برای نظام بازیابی اطلاعات وجود ندارد. در این مقاله، ما یک چهارچوب نظری با عنوان «اِن اِل پی آی آر» را با هدف پشتیبانی و عمومی‌سازی برنامه پردازش زبان طبیعی در بازیابی اطلاعات ارائه می‌کنیم.

چهارچوب «اِن اِل پی آی آر» بر این فرض استوار است که فاصله بازنمونی بین پرسش و مدارک وجود دارد. برای کاهش این فاصله و دستیابی به عملکرد بهتر نظام بازیابی اطلاعات، این چهارچوب برخی فناوری‌های پردازش زبان طبیعی را با سطوح مختلف فرایند بازیابی اطلاعات ادغام می‌کند. این چهارچوب رویکردهای مختلف این ادغام را در پنج گروه رویکرد مستقیم، گسترشی، استخراجی، تحولی و اتحادی قرار می‌دهد.

بقیه مقاله به شرح زیر سازماندهی شده است. ما ابتدا مشکلات موجود در بازیابی اطلاعات سنتی مبتنی بر کلیدواژه را بررسی می‌کنیم و نقش پردازش زبان طبیعی را در پرداختن به برخی از این مشکلات توصیف می‌کنیم. سپس، یک چهارچوب نظری به نام «اِن اِل پی آی آر» پیشنهاد می‌کنیم که بتواند از برنامه‌های گسترده پردازش زبان طبیعی در بازیابی اطلاعات پشتیبانی کند و یک طبقه‌بندی منظم از رویکردها برای تحقق این اهداف ارائه دهد. در مرحله بعد، برخی از فنون دقیق پردازش زبان طبیعی برای نشان دادن چهارچوب استفاده می‌شود. سرانجام، به طور خلاصه درباره نظرات مختلفی راجع به رابطه بین پردازش زبان طبیعی و بازیابی اطلاعات و کاربردهای بالقوه چهارچوب «اِن اِل پی آی آر» بحث خواهیم کرد.

مشکلات بازیابی اطلاعات سنتی مبتنی بر کلیدواژه

یک نظام بازیابی اطلاعات کارا باید به سرعت بتواند اطلاعات مرتبط را شناسایی کند تا نیازهای کاربر را برآورده کند و قطعات بی‌ربط را نادیده گیرد. تاکنون افراد بیشتر از سه روش شامل فنون مبتنی بر دانش، روش‌های احتمالی و فنون یادگیری ماشین برای دستیابی به بازیابی اطلاعات «هوشمند» استفاده کرده‌اند.

فنون دانش‌بنیان سعی در جذب دانش متخصصان و جویندگان اطلاعات، استراتژی‌های جستجو و اکتشاف پذیری پالایش پرس‌وجو در نظام‌های بازیابی اطلاعات هستند (چن و دهر[1]، ۱۹۹۱). پایگاه دانش، توانایی تجزیه و تحلیل محتوا و استدلال را در اختیار کاربران قرار می‌دهد و نظام بازیابی اطلاعات را قادر می‌سازد تا به طیف وسیعی از سؤالات پاسخ دهد. به هرحال، در واقعیت، حتی برای یک حوزه کوچک نیز، ایجاد پایگاه دانش با تمام قواعد ممکن، امکان‌پذیر نیست. علاوه بر این، تلاش‌های چشمگیر بشری اغلب برای کسب دانش و نگهداری از پایگاه دانش مورد نیاز است. طی دهه‌های گذشته، بازیابی اطلاعات با استفاده از روش‌های احتمالی توجه قابل‌توجهی را به خود جلب کرده است (بوکستین و سوانسون[2]، ۱۹۷۵؛ مارون و کوهن[3]، ۱۹۶۰)؛ زیرا گفتگوی شفاهی بین کاربر و نظام وجود ندارد تا به کاربر گفته شود که مدارک موجود در یک مجموعه مرتبط هستند یا نیستند، بلکه کاربر مجبور است از دانش ناقص خود برای تعیین ربط مدارک استفاده کند، خواه مدارک ارائه‌شده مرتبط به پرس‌وجو باشد یا نه. یک روش معقول برای نشان دادن سطح اطمینان یک تصمیم، تخمین احتمال ربط آن است. این رویکرد مبتنی بر فرضیات احتمالی است که از قبل در مورد توزیع عناصر در مدارک مربوط و نامربوط ایجاد شده است. باوجود الحاقات مختلف، روش‌شناسی احتمالی هنوز مستلزم فرض مستقل در مورد اصطلاحات است و باید با دشواری‌های تخمین صحیح رخداد اصطلاحات سازگار شود.

فنون یادگیری ماشین در بازیابی اطلاعات نیازمند کسب دانش از مجموعه‌های آموزشی یا مثال‌هایی[4] هستند. فنونی که اغلب مورد استفاده قرار می‌گیرند شامل الگوریتم‌های یادگیری استقرایی، یادگیری نمادین و الگوریتم‌های ژنتیکی هستند. برای استفاده از فناوری شبکه عصبی در ساختن مدل‌های کاربر، کاربر مجموعه اولیه مقالات خبری را بازیابی و آن‌ها را با توجه به علاقه خود به عنوان مرتبط یا نامرتبط علامت‌گذاری می‌کند. سپس این نمونه‌های مثبت و منفی برای آموزش به شبکه عصبی به منظور بازیابی به عنوان ویژگی‌های متداول در مقالات مرتبط مورد استفاده قرار می‌گیرد (جنینگز و هیگن چی[5]، ۱۹۹۲). در یک الگوریتم ژنتیکی سازگار با بازیابی اطلاعات (چن ، ۱۹۹۵)، یک ژن و یک کروموزوم فردی به ترتیب یک کلمه کلیدی و یک مدرک را نشان می‌دهد. مجموعه اولیه شامل مدارک انتخاب شده توسط کاربر است. اگر کاربر، منابع مرتبط را فراهم آورد، متوسط مناسبت برای مجموعه کامل مدارک بیشتر خواهد بود. نظام، تجزیه و تحلیل نتایج جستجوی میانی کاربران را تسهیل و سایر مدارک بالقوه مرتبط را نیز پیشنهاد می‌کند.

1. Chen & Dhar
2. Bookstein & Swangon
3. Maron & Kuhns
4. Training corpus or examples
5. Jennings & Higuchi

محدودیت روش یادگیری ماشین این است که به شدت به مثال‌های آموزشی بستگی دارد که چه بسا ممکن است، نتایج مغرضانه باشد.

در حال حاضر، اکثر موتورهای کاوش از جستجوی مبتنی بر کلمات کلیدی سنتی استفاده می‌کنند که توسط پرس‌وجوی بولی تقویت شده است. در این روش، هر مدرک با یک بردار اصطلاحات وزنی که می‌توانند کلمات یا عبارات باشند، نمایش داده می‌شود. این طرح استاندارد وزن دهی «تی اف در آی دی اف[1]» است (سالتون[2]، ۱۹۸۹). «تی اف»، شکل کوتاه شده فراوانی یک اصطلاح است که به وقوع فرکانس اصطلاح «تی» در یک مدرک اشاره دارد، و «آی دی اف» معکوس فرکانس مدارک است که از طریق فرمول لگاریتم تخمین زده می‌شود که به تعداد کل مدارک موجود در یک مجموعه و تعداد مدارک دارای اصطلاح «تی» اشاره دارد. پرس‌وجو با اصطلاح‌هایی از اصطلاحات نمایه‌ای به همراه ترکیبی از پیوندهای منطقی معمول (منطق بولی) نظیر «و»، «یا» و «نه» اظهار می‌شود. یک موتور کاوش بولی، مدارکی را برای یک پرسش درست (مرتبط) دانسته و بازیابی خواهد کرد که این مدارک دارای اصطلاح‌های موجود در پرسش هستند. یک پرس‌وجوی بولی مبتنی بر کلمه کلیدی خوشایند به نظر می‌رسد، اما برخی از مشکلات ابهام ذاتی را دارد که منجر به کارایی ضیف نظام بازیابی اطلاعات می‌شود که در ادامه به این مشکلات اشاره شده است.

۱. مترادف‌ها[3]: یک مفهوم می‌تواند با اصطلاح‌های مختلفی که مترادف خوانده می‌شود؛ اظهار شود. بدون بررسی مترادف در یک نظام مطابقت دقیق، ممکن است باعث نادیده گرفتن شدن مدارکی شود که فقط حاوی مترادف کلمات کلیدی هستند، نه خود کلمات کلیدی.

۲. چندمعنایی‌ها[4]: یک کلمه می‌تواند چندین حس داشته باشد و ممکن است در صورتی که واژه خارج از بافت (متن) بررسی شود، ابهام ایجاد کند. حتی اگر کلمات کلیدی یک پرس‌وجو در یک مدرک نیز موجود باشد، ممکن است معنای نامرتبط داشته باشد. مثلاً واژه «شیر» دارای معانی متعدی نظیر شیر حیوان، شیر خوردنی و شیر آب است؛

۳. اولویت کلمه کلیدی[5]: وقتی کاربر چندین کلمه کلیدی را در یک پرس‌وجو سیاهه می‌کند، ممکن است ترجیحات واضح اما نامشخص در بین آن‌ها داشته باشد؛

۴. ناتوانی در تشخیص نیاز: در شرایط خاص، کاربر ممکن است احساس کند که فرمول‌بندی پرس‌وجو به‌گونه‌ای که به وضوح علایق وی را بیان‌کنند، مشکل است. به عنوان مثال، تفکیک بین علاقه به این که «بانک اطلاعاتی چیست» و «چرا به بانک اطلاعاتی نیازمندیم»، چالش‌برانگیز است؛

1. Term frequency–inverse document frequency
2. Salton
3. Synonyms
4. Polysemy
5. Keyword preference

۵. ریزش کاذب[1]: بسیاری از موتورهای کاوش اطلاعات را بدون درک معنای آن بازیابی می‌کنند. هنگام مقایسه کلمات کلیدی در یک پرس‌وجو با مدارک موجود، ممکن است فقط به دنبال مطابقت دقیق باشد؛

۶. بافت[2]: هر کلمه در یک مدرک دارای بافت خاصی است که معنای آن را تعیین می‌کند، اما چنین اطلاعاتی در یک پرس‌وجو مبتنی بر کلمه کلیدی در دسترس نیست؛

۷. سیاهه طولانی از اصطلاح‌های نمایه‌ای: برخی از کلمات منفرد، اصطلاح‌های نمایه‌ای خوبی فقط در عبارات هستند. بنابراین، هنگامی‌که یک نظام اطلاعاتی می‌کوشد شرایط نمایه‌سازی را از یک مدرک به دست آورد، باید هر ترکیب ممکن از کلمات همسایه را به عنوان نامزد بالقوه در نظر بگیرد؛ زیرا نظام با هر کلمه به عنوان یک نماد رفتار می‌کند. در نتیجه، سیاهه‌های طولانی از اصطلاح‌های نمایه‌ای ایجاد می‌شود که اصلاً تعداد زیادی از آن‌ها اصطلاح‌های مهمی نیستند. این منجر به یک فضای ذخیره‌سازی بزرگ نمایه‌ای و جستجوی ناکارآمد می‌شود.

پژوهشگران مطالعات گسترده‌ای در مورد چگونگی بهبود بازیابی مبتنی بر کلمه کلیدی انجام داده‌اند. اتخاذ فناوری پردازش زبان طبیعی به صورت گسترده بحث شده است. تا به امروز، انواع فنون پردازش زبان طبیعی برای رفع مشکلات ذکرشده در بالا به کار گرفته شده است. برخی از روش‌های گسترش پرس‌وجو، مانند استفاده از پروفایل کاربر و بازخورد ربط و رویکردهای بهبود عملکرد نظام‌های بازیابی، بردار استاندارد–فضایی، مانند وزن‌گیری کوتاه‌مدت نیز دارای پتانسیل هستند. در این مقاله به فنون پردازش زبان طبیعی پرداخته شده است.

چهارچوبی برای به کارگیری پردازش زبان طبیعی در بازیابی اطلاعات

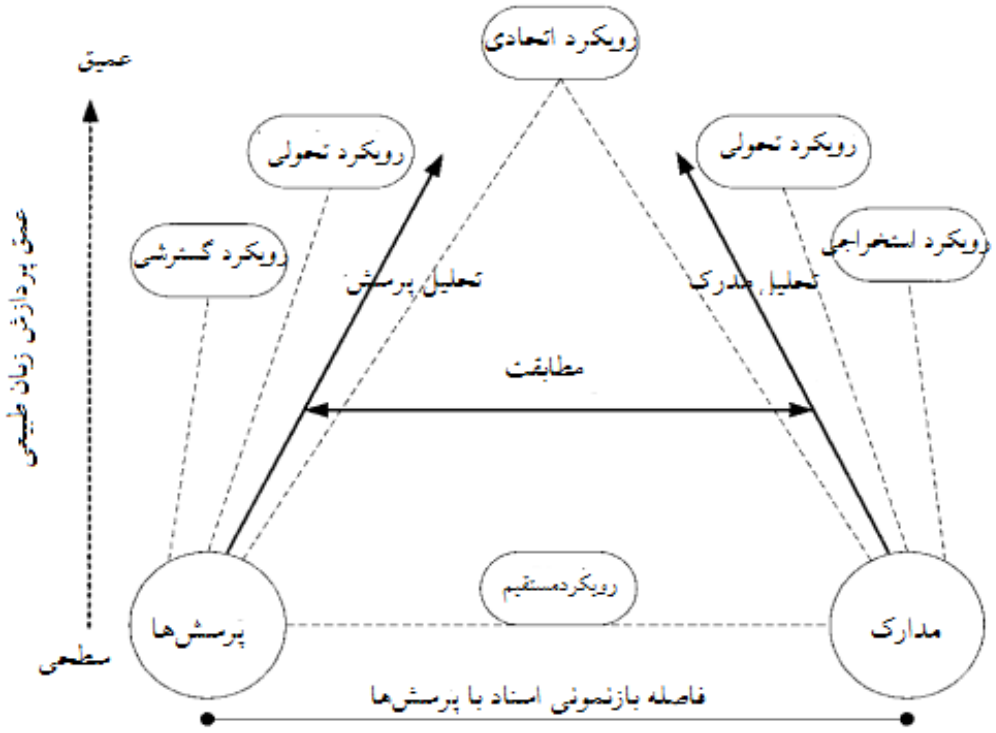
پژوهش‌های پردازش زبان طبیعی به این سؤال مهم پاسخ می‌دهند که چگونه مردم معنای یک جمله یا مدرکی را درک می‌کنند؛ با دانستن این که چه کسی چه کاری را انجام داده، برای چه کسی انجام داده، چه چیزی را انجام داده است. اگرچه کلمات منفرد بلوک‌های معنا هستند، اما واژه‌ها دارای رابطه با جمله، مدرک و بافت نیز هستند. بافت یعنی، آنچه در حال حاضر در مورد جهان می‌دانیم که معنای واقعی یک متن را منتقل می‌کند. بنابراین، پردازش زبان طبیعی به عنوان ابزاری قدرتمند برای تسهیل در درک و تجزیه و تحلیل متن شناخته شده است که مبتنی بر این پیش‌فرض است که اطلاعات معنایی انتقادی[3] را در مدارک کشف می‌کند و شمارش ساده کلمات نمی‌تواند موجب درک معنای آن شود. پژوهش‌های پردازش زبان طبیعی در سال‌های اخیر به خصوص در زمینه‌های بازیابی اطلاعات مبتنی بر متن، استخراج اطلاعات، خلاصه‌سازی و ترجمه چندزبانه دستاوردهای چشمگیری داشته است. هدف پردازش زبان طبیعی در بازیابی

1. False drops
2. Context
3. Critical semantic information

اطلاعات، بهبود درک و بازنمایی متون با ایجاد مدل‌های محاسباتی از زبان است تا افراد بتوانند برنامه‌های رایانه‌ای را به‌گونه‌ای بنویسند که آن‌ها قادر به انجام کارهای مختلفی از جمله زبان طبیعی باشند.

مزایای پردازش زبان طبیعی در بازیابی اطلاعات به طور گسترده بحث شده است. راو و ژاکوبز[1] (۱۹۸۹) اظهار داشتند که پردازش زبان طبیعی می‌تواند در سه زمینه عمومی بازیابی اطلاعات شامل رابط زبان طبیعی، پردازش متن و کسب دانش به کار گرفته شود. لوئیس و جونز[2] (۱۹۹۶) نقش پردازش زبان طبیعی را در جستجوی متن کامل بررسی کردند که در آن می‌تواند به سه نوع چیز کمک کند: (الف) کلمات، عبارات و جملات ضبط‌شده توسط دستگاه پردازش متن؛ (ب) ساختار «طبقه‌بندی» بر روی یک مدرک به عنوان یک کل که نشان‌دهنده روابط پارادایمی بین اصطلاحات است و موجب جایگزینی واژگان کنترل‌شده می‌شود (پ) مکانیسم‌های مبتنی بر پردازش زبان طبیعی برای جستجو و مطابقت. هوی[3] (۱۹۹۸) در مورد کاربرد زبان طبیعی در بازیابی اطلاعات و خلاصه‌سازی بحث کرده است، از نظر وی در هر دو سطح و بالاتر از سطح کلمه قابل‌استفاده است. به ویژه، در سطح کلمه، تکواژشناسی در چهار زمینه الگوریتم‌های ریشه‌یابی[4]، توسعه فرهنگ لغات و اصطلاح‌های قابل‌خواندن با ماشین، نمایه‌سازی حس کلمه و ابهام‌زدایی حس کلمه (وی اس دی[5]) قابل‌استفاده است. این حوزه از سطح کلمه که در بالا یاد شد، شامل بازنمایی مفهومی، نمایه‌سازی نحوی و معنایی و چکیده‌نویسی خودکار است. در حالت کلی، ابهام‌زدایی حس کلمه در بازیابی اطلاعات ارزشمند شناخته می‌شود؛ اگرچه نتایج آن قانع‌کننده نیست (ساندروسن[6]، ۱۹۹۴؛ ویلکس[7]، ۱۹۹۹) ویلکس استفاده از موقعیت را پیشنهاد داد. به عنوان یکی از کارکردهای اصلی در پروژه «ان جی آی آر[8]»، یک مکانیسم پرسش و پاسخ عمومی نیاز به یک نظام بازیابی اطلاعات دارد تا بتواند به صورت خودکار پاسخ خود را از اطلاعات تکه تکه یا حتی از اجزا ناهماهنگ توزیع‌شده در بین مدارک یا منابع متعدد جمع‌آوری کند (استارلایزکوئسکی، استاین، وایز و باقا،[9] ۲۰۰۰).

1. Rau and Jacobs
2. Lewis and Jones
3. Hui
4. Stemming
5. word sense disambiguation
6. Sanderson
7. Wilks
8. NGIR
9. Strzalkowski, Stein, Wise, & Bagga



شکل ۱. چهارچوب به‌کارگیری پردازش زبان طبیعی در بازیابی اطلاعات

بنابراین، پردازش زبان در سطح تحلیل معنا و تحلیل گفتمان نیز مورد نیاز خواهد بود. تجربه «ان جی آی آر» نشان داد که برای دستیابی به عملکرد رضایت‌بخش به‌گونه‌ای که نیازهای اطلاعاتی پیشرفته را پشتیبانی کند، لازم است تا قابلیت‌های پردازش زبان طبیعی یکپارچه شوند. فناوری پردازش زبان طبیعی تقریباً در تمام مراحل بازیابی اطلاعات نظیر پردازش مدارک، پردازش پرس‌وجو و تطبیق قابل‌استفاده است. در طول مرحله پردازش مدارک، فنون قدرتمند پردازش زبان طبیعی می‌توانند به بازنمایی بهتر مدارک متنی برای نمایه‌سازی و جستجوی اهداف کمک کنند. علاوه بر بهبود تحلیل و گسترش پرسش کاربران، پردازش زبان طبیعی می‌تواند در پردازش پرس‌وجو به درک دقیق‌تر نیازهای اطلاعاتی کمک کند. در طی مرحله تطبیق، پردازش زبان طبیعی با اجازه دادن به تطبیق و رتبه‌بندی در سطح ساختار و معنی به جای سطح کلمه، انعطاف‌پذیری و دقت بالاتری را ارائه می‌دهد. علاوه بر این، امکان جمع‌آوری پاسخ‌ها از تکه‌های جداگانه اطلاعاتی فراهم می‌شود.

به طور خلاصه، پژوهش‌ها نشان داده است که پردازش زبان طبیعی می‌تواند عملکرد نظام‌های بازیابی اطلاعات را در جنبه‌های مختلف بهبود بخشد. با وجود این، ما دو مشکل را شناسایی کردیم که بندرت بحث شده است. مشکل اول این است که پژوهش و توسعه مؤلفه‌های پردازش زبان طبیعی در یک نظام بازیابی اطلاعات فاقد پشتیبانی از یک چهارچوب است و منجر به درک ناقص و به ظاهر متناقض از نقش پردازش زبان طبیعی در بازیابی اطلاعات می‌شود.

دوم این‌که پردازش زبان طبیعی و بازیابی اطلاعات کاملاً متصل به هم هستند؛ زیرا بازیابی اطلاعات تنها یکی از کاربردهای پردازش زبان طبیعی است و پردازش زبان طبیعی صرفاً رویکردی برای تقویت نظام‌های بازیابی اطلاعات است. ادغام پردازش زبان طبیعی و بازیابی اطلاعات در یک چهارچوب منسجم، بینشی از رابطه بین آن‌ها را ارائه می‌دهد. در بخش بعدی چهارچوبی با عنوان «اِن اِل پی آی آر» را ارائه خواهیم کرد که برای پرداختن به دو موضوع فوق پیشنهادشده است.

چهارچوب «اِن اِل پی آی آر»

چهارچوب «اِن اِل پی آی آر» برای پشتیبانی و گسترش پژوهش و توسعه ترکیبی از پردازش زبان طبیعی و بازیابی اطلاعات ارائه شده است. همان‌طور که در شکل ۱ نشان داد شد، چهارچوب در اصل از سه نوع مؤلفه تشکیل‌شده است: پرسش و مدارک بازنمون شده در یک دایره؛ تجزیه و تحلیل پرسش، تجزیه و تحلیل مدارک و تطبیق آن با فلش‌های مستقیم و رویکردهای مختلف مربوط به پرسش، مدارک و هر دو توسط خطوط سایه‌روشن مشخص‌شده‌اند. عمق پردازش زبان طبیعی به تدریج از پایین به بالا در شکل ۱ افزایش می‌یابد. به عنوان مثال، یک پرس‌وجو یا مدرک ممکن است از تکواژشناسی گرفته تا تحلیل نحوی و سپس تحلیل معنا تحت پردازش زبان طبیعی قرار بگیرد. ما در ادامه سه نوع مؤلفه را به صورت جداگانه معرفی می‌کنیم.

در اصل، بازیابی اطلاعات فرایندی برای مطابقت پرسش با مدارک است، اما درخواست کاربر با مدارکی که می‌خواهد به هر طریقی بازیابی کند، متمایز است. یک تفاوت این است که یک پرسش به طور قابل توجهی مختصر است که معمولاً از چندین کلمه کلیدی یا تقاضایی به زبان طبیعی تشکیل شده است. حتی ممکن است یک پرس‌وجو به زبانی متفاوت از یک مدرک بیان شود. بنابراین، یک فاصله باز نمونی بین پرسش و مدارک در یک نظام بازیابی اطلاعات وجود دارد. چنین فاصله‌ای می‌تواند ناشی از تفاوت در کاربرد کلمه، ساختار نحوی، ساختار معنایی، گفتمان، کاربردشناسی، معنی و خود زبان و همچنین از ابهامات در تمام این سطوح باشد. برای کاهش فاصله و ابهامات بازنمونی، می‌توان پایه زبانی مؤثری را با استفاده از پردازش زبان طبیعی برای مقایسه پرسش و مدارک ایجاد کرد. سه فلش مستقیم نزدیک به مرکز شکل ۱ نمایانگر فرایندهای سنتی بازیابی اطلاعات شامل تحلیل پرسش، مدارک و مطابقت بین آن‌هاست. تجزیه و تحلیل‌گر مدارک، یک مدرک را به‌گونه‌ای پردازش می‌کند که می‌تواند توسط نظام بازیابی ذخیره و استفاده شود. به طور مشابه، یک پرس‌وجو باید از طریق تجزیه و تحلیل‌گر پرس‌وجو پردازش، و بازنمونی از آن تولید شود تا نظام بازیابی اطلاعات بتواند آن را با بازنمون یک مدرک مقایسه کند. در فرایند تطبیق، یک پرس‌وجوی تفسیر شده با مدارک از پیش پردازش شده مقایسه می‌شود تا ربط یک مدرک با پرس‌وجو مشخص شود. در هر یک از این سه فرآیند، می‌توان از سطوح مختلف استفاده کرد. همان‌طور که در شکل ۱ مشاهده می‌شود، نزدیکی تجزیه و تحلیل پرس و جو و مدارک به فلش، نیاز عمیق‌تر به پردازش زبان طبیعی را نشان می‌دهد. بر این اساس، مطابقت در سطح زبانی عمیق‌تر انجام خواهد

شد. فلش تطبیق دو سر لازم نیست که افقی باشد، زیرا تطبیق می‌تواند مسئولیت بیشتری را برای استفاده بیشینه سطوح مختلف پردازش زبان طبیعی در پرسش و مدارک بر عهده گیرد.

هم پرسش و هم مدارک در مجموعه رویکردهای چهارچوب با همدیگر مرتبط هستند. هر رویکرد یک سری بازنمودهای واسطه از پرسش یا مدارک ایجاد می‌کند که توسط یک خط نقطه‌دار نشان داده‌شده است. نتیجه استفاده از پردازش زبان طبیعی را می‌توان به عنوان حرکت پرسش یا مدرک مشاهده کرد که در طول یک جفت خط به یکدیگر نزدیک‌تر می‌شوند؛ بنابراین تطابق بین پرسش و مدارک ساده‌تر و دقیق‌تر می‌شود. همان‌گونه که در شکل ۱ مشاهده می‌شود، هرچه نقاط انتهای بالای هر دو خط پرسش و مدارک نزدیک‌تر شوند، روند تطبیق به تلاش کمتری نیاز دارد. به عبارت‌دیگر، پردازش زبان طبیعی بازنمایی‌های پرسش و مدارک را به هریک نزدیک‌تر می‌کند و پیچیدگی و ابهامات روند تطبیق را کاهش می‌دهد. در همین راستا، رویکردهای « اِن اِل پی آی آر» نشان می‌دهند که چگونه پردازش زبان طبیعی و بازیابی اطلاعات به شدت در چهارچوب یکپارچه شده‌اند. در ادامه هر یک از رویکردها با جزئیات بیشتر شرح داده شده است.

رویکردهای «اِن اِل پی آی آر»

کلی‌سازی رویکردهای «اِن اِل پی آی آر » برای چهارچوب طراحی‌شده «اِن اِل پی آی آر » به عنوان راهنمایی برای اقتباس از فنون مختلف پردازش زبان طبیعی و منابع در بازیابی اطلاعات قابل انتقاد است. این رویکردها با توجه به عمق پردازش زبان طبیعی درگیر در فرآیندهای بازیابی اطلاعات به پنج دسته طبقه‌بندی می‌شوند.

رویکرد مستقیم[1]

این رویکرد اساساً مبتنی بر مقایسه کلمات بین پرس‌وجو و مدرک است. این امر بسیار شبیه به تطابق با کلمات کلیدی است، به جز این که ممکن است، برخی از سطوح پردازش زبان طبیعی نظیر ریشه‌یابی و بن‌واژه‌یابی[2] به کار گرفته شوند. همان‌طور که گفته شد، تطابق کلمات کلیدی به صورت ساده به دلیل ابهامات گوناگون موجود در بازیابی اطلاعات می‌تواند مشکل‌ساز باشد. در نتیجه، فاصله بازنمایی بین پرسش و مدارک بدون تغییر باقی‌مانده است و معمولاً نتایج بازیابی به جز مطابق دقیق غیرقابل‌قبول است. برای تولید نتایج جستجوی بهتر، تجزیه و تحلیل کامل‌تری از پرسش و مدارک مورد نیاز است.

1. Direct Approach
2. Stemming and lemmatization

رویکرد گسترشی[1]

از آنجا که مشکل اصلی رویکرد مستقیم عدم وجود اطلاعات واژگانی–معنایی درباره پرسش‌ها و مدارک است، محققان در تلاش هستند تا روش‌هایی برای گرفتن اطلاعات پیدا کنند. آن‌ها نظام‌های بازیابی اطلاعات گسترشی را توسعه داده‌اند که در آن از اصطلاح‌نامه‌ها برای گسترش پرس‌وجو استفاده می‌شود. نتایج آمیخته[2] از چنین تلاش‌هایی گزارش‌شده است، تا حدی به این دلیل که اصطلاح‌نامه‌ها عمومی هستند. برای افزایش دقت در گسترش پرس‌وجو در یک دامنه ، اخیراً متخصصان بازیابی اطلاعات به هستی‌شناسی دامنه توجه زیادی کرده‌اند. با این‌حال، اصل اساسی تقویت نظام‌های بازیابی اطلاعات با یک هستی‌شناسی، مشابه اصطلاح‌نامه است.

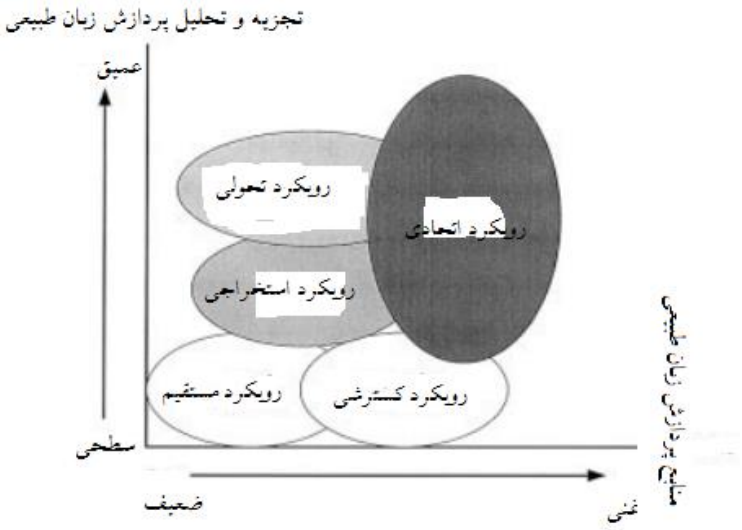
رویکرد استخراجی[3]

رویکرد استخراجی با فناوری استخراج اطلاعات مرتبط است. هدف آن استخراج انواع خاصی از اطلاعات قابل‌توجه برای نیازهای کاربر از یک مدرک و یا پرس‌وجو است. در همین حال، بخش‌هایی از متون که مرتبط هستند، نادیده گرفته می‌شوند. روش‌های عمده از انواع پردازش زبان طبیعی و فن‌های آماری برای استخراج انواع حقایق از پیش تعیین‌شده، مانند شخص، سازمان، مکان‌ها و مبالغ پولی از متن استفاده می‌کند. علاوه بر این، واژه‌نامه‌ها و فرهنگ لغات به زبان طبیعی به شدت درگیر می‌شوند. اگرچه از نظر ماهوی، بازیابی اطلاعات سنتی، آماری در نظر گرفته می‌شود، ویلکس (۱۹۹۹) دریافت که تمایز بین بازیابی اطلاعات و فناوری استخراج اطلاعات با رشد روش‌های یادگیری در فناوری استخراج اطلاعات در سال‌های اخیر از بین رفته است.

رویکرد تحولی[4]

این روش با استفاده از پردازش زبان طبیعی نسبتاً عمیق، پرسش و مدرک را به نوعی بازنمایی متوسط تبدیل می‌کند. این در تبدیل مواد اولیه، پالایش و کاهش آن به عناصر متن اصلی نقش مهمی دارد.

1. Expansion Approach
2. Mixed results
3. Extraction Approach
4. Transformation Approach



شکل۲. گروه‌بندی رویکردهای «اِن اِل پی آی آر»

رویکرد تحولی به بافت کلمات (یا عبارات) در یک پرس‌وجو یا یک مدرک توجه می‌کند. بازنمون واسطه می‌تواند مبتنی بر تحلیل نحو یا تحلیل معنا باشد. فرایند تطبیق شباهت‌های بین پرسش و مدرک بر اساس کلمات یا عبارات موجود در آن‌ها و همچنین نقش‌هایی که آن‌ها در متن دارند، مشخص می‌شود.

رویکرد اتحادی[1]

تحلیل عمل‌گرایانه و گفتمانی در این رویکرد در نظر گرفته شده است. رویکرد اتحادی ساختار انواع مختلفی از مدارک را مورد بررسی قرار داده و تقاضای بدنه دانش آن‌سوی دانش جهانی یعنی فراتر از محتوای پرسش و مدرک را مورد نظر قرار می‌دهد که این امر برای استنتاج و ایجاد حس در نظام بازیابی اطلاعات لازم است تا بهترین جواب‌ها را به یک پرسش بازگرداند. یکی از راه‌های اجرای رویکرد اتحادی، طراحی یک زبان است که می‌تواند تمام اطلاعات مفید برای بازیابی اطلاعات را در گفتمان‌ها ، زبان‌ها و کاربردهای مختلف ثبت کند. پردازش زبان طبیعی می‌تواند از ساختارهای قابل پیش‌بینی برای شناسایی نقش تکه‌های اطلاعاتی در یک مدرک استفاده کند. اگر ما می‌توانیم بازنمون متحد پیدا کنیم که هر پرسش و مدرک می‌تواند به آن تبدیل شود، به تلاش کم‌تری برای مطابقت بین پرسش و مدارک نیاز است. در شکل ۱ ، اثر رویکرد اتحادی توسط دو خط برگرفته از پرسش و مدرک نشان داده شده است که در نهایت به هم متصل می‌شوند. به عبارت‌دیگر، در این رویکرد به اقدامات مطابقتی کم‌تری نیاز است؛ زیرا فاصله بازنمایی بین پرسش و مدرک به حداقل کاهش یافته است. با این‌حال، در عمل، هنوز هم یک کار بسیار چالش‌برانگیز است که محققان دانش

1. Uniform Approach

جهانی مورد نیاز خود را در هنگام نیاز به آن تعریف کنند. رویکردهای مختلف «اِن اِل پی آی آر» متضاد هم نیستند، بلکه مکمل یکدیگرند. افزون بر این، کاربرد پردازش زبان طبیعی در بازیابی اطلاعات نه تنها در مدل‌های تحلیل، بلکه در منابع زبانی مانند اصطلاح‌نامه‌ها منعکس می‌شود. از نظر تحلیل پردازش زبان طبیعی و منابع به کار رفته در بازیابی اطلاعات، روابط بین پنج رویکرد در شکل ۲ نشان داده شده است. برخی از رویکردها در عمل ترکیب شده‌اند. برخی از آن‌ها برای پردازش مدارک در مقایسه با پرسش‌ها مناسب‌ترند و بالعکس. به عنوان مثال، فناوری‌های چکیده‌نویسی و خلاصه کردن متون معمولاً مرتبط با مدارک است، همان‌طور که در نمونه اولیه «ان جی آی آر» نشان داده شده است، کاربران می‌توانند به صورت تعاملی اطلاعات خلاصه‌ای را در مورد موضوعات مورد علاقه گردآوری کنند.

فنون کاربردپذیر پردازش زبان طبیعی در چهارچوب «اِن اِل پی آی آر»

برای اعتبارسنجی چهارچوب «اِن اِل پی آی آر» با جزئیات خاص، در ادامه برخی از فنون پردازش زبان طبیعی در هر یک از رویکردهای «اِن اِل پی آی آر» انتخاب و به طور خلاصه شرح داده شده است.

رویکرد مستقیم

جدا کردن واژه‌ها و ریشه‌یابی[1]: در پایین‌ترین سطح، نظام‌های بازیابی اطلاعات باید متن را به منظور جستجو به واحدهای اساسی آن بشکنند. جدا کردن واژه‌ها در زبان انگلیسی ساده است، اما در زبان‌های دیگر مانند آلمانی یا چینی یک چالش بسیار بزرگی است. افزون بر این، بسیاری از نظام‌های بازیابی اطلاعات به منظور از بین بردن انتهای تورمی متداول تکواژشناسی واژها، فرایند ریشه‌یابی را نیز به کار می‌گیرند. یک واژه ممکن است در اشکال مختلف ظاهر شود. به عنوان مثال خواندن، خوانش، خوانا، خواندم همه فرم اصلی و معنای مشابه «خوان» را به اشتراک می‌گذارند. هنگامی‌که پرس‌وجوهای بولی کوتاه هستند، ممکن است که تغییرات تکواژ شناختی به دلیل انتخاب‌های مختلف واژه‌ها باعث حذف مدارک مربوطه شود. امروزه، بیشتر موتورهای کاوش به طور خودکار ریشه کلمات پرس‌وجو را از اشکال مفرد و جمع می‌گیرند. تعدادی از الگوریتم‌های ریشه‌یابی بر اساس قواعد زبانی طی سال‌ها تدوین‌شده است.

متوقف کردن ادات سخن[2]: استفاده از واژگان کنترل نشده در بازیابی اطلاعات برخی از مشکلات را موجب می‌شود. بسیاری از مطالعات نشان داده‌اند که ۲۵۰ تا ۳۰۰ کلمه متداول در انگلیسی ممکن است ۵۰ درصد یا بیشتر از هر متن را به خود اختصاص دهد. این کلمات به خودی خود بار معنایی کمی دارند و تمایل به کاهش تأثیر اختلاف فرکانس در بین کلمات کمتر متداول دارند. علاوه بر این، آن‌ها باعث مقدار زیادی پردازش غیرضروری می‌شوند. بنابراین، این غیرمعمول نیست که یک سیاهه کلمات توقف تهیه کنیم. در نتیجه، کلمات متوقف اغلب در هنگام تحلیل سؤالات مربوط به زبان

1. Tokenization and stemming
2. Stop-POSing

طبیعی و مدارک نادیده گرفته می‌شوند. با الهام از فن توقف جمله، ژو ، بوکر، و ژانگ[1] (۲۰۰۲) مفهوم ایست–ادات سخن[2] را پیشنهاد کردند. درحالی‌که یک متن در حال پردازش است، بخشی از گفتار یا سخن[3] هر کدام از کلمات در برابر سیاهه ایست – ادات سخن بررسی می‌شود. اگر در این سیاهه قرار داشته باشد، کلمه مربوطه بدون پردازش اضافی پالایش می‌شود. معمولاً اسم‌ها و صفت‌ها در بازیابی اطلاعات حفظ می‌شوند، درحالی که برخی از بخش‌های گفتاری دیگر مطابق با شرایط فردی حذف می‌شوند. آرامپاتیز، واید ، کوستر و بومم[4] (۲۰۰۰) گزارش دادند که یک مجموعه فهرست‌بندی اصلی مبتنی بر کلیدواژه می‌تواند اسم و صفت را فقط بدون آسیب رساندن به بازیابی، یا حتی اندک بهبود در آن حفظ کند.

رویکرد گسترشی[5]

بدیهی است که یک مدرک یک سؤال را پاسخ می‌دهد یا نمی‌دهد. اگر برای نشان دادن محتوای یک مدرک و یک پرس‌وجو به طور جداگانه از مدل فضای برداری استفاده کنیم، شباهت بین آن‌ها با تابع کسینوس دو بردار قابل‌اندازه‌گیری است (سالتون، ۱۹۸۹). مدارک همسان بر اساس فراوانی هم‌زمانی کلمات کلیدی مشخص شده در پرس‌وجو رتبه‌بندی می‌شوند. با این‌حال، یک مدرک ممکن است با یک پرس‌وجو مطابقت داشته باشد، بدون این‌که بین آن‌ها کلمات مشترک وجود داشته باشد. گسترش جستجوی کاربر با مترادف‌ها یا اصطلاحات مرتبط می‌تواند عملکرد جستجوی نظام بازیابی مبتنی بر متن را بهبود بخشد. به عنوان مثال، اگر درخواست کاربر حاوی عبارت «بازیابی اطلاعات» باشد، می‌توان یک عبارت معادل «بازیابی اطلاعات» به پرسش اولیه اضافه کرد. منابع معمول برای گسترش یاد شده شامل اصطلاح‌نامه و هستی‌شناسی است.

اصطلاح‌نامه[6]

رویکرد گسترش پرس‌وجو، اضافه کردن اصطلاحات مشابه یا مرتبط با آن به پرسش اصلی، استفاده از یک اصطلاح‌نامه آنلاین است تا اصطلاحات پرس‌وجو به عباراتی تبدیل شوند که با نمایه مدرک مطابقت دارند. به طورکلی دو نوع منابع متنی برای دستیابی به داده‌های واژگانی وجود دارد: فرهنگ لغت‌های قابل‌خواندن با ماشین و شرکت متن. تقریباً تمام اصطلاح‌نامه‌های تهیه‌شده به صورت خودکار، مبتنی بر هم رخدادی آماری انواع کلمات هستند که با تجزیه و تحلیل نحوی متون ایجاد شده‌اند. ورد نت[7] یک نظام مرجع واژگانی آنلاین است که توسط دانشگاه پرینستون برای به دست

1. Zhou, Booker, and Zhang
2. Stop-POSing
3. Part-of-speech
4. Arampatzis, Weide, Koster, and Bomme
5. Expansion Approach
6. Thesauri
7. WordNet

آوردن مجموعه‌های مترادف یک کلمه معین که به طور گسترده مورد استفاده قرار گرفته است، ساخته شده است (میلر، ۱۹۹۹). به عنوان مثال، اسمیتون[1] (۱۹۹۹) از رابطه بین مجموعه‌های مترادف در وردنت برای به دست آوردن شباهت‌های معنایی بین اسم‌ها در بازیابی اطلاعات استفاده کرد که محتوای اطلاعات اولین مجموعه مترادف بود که دو مجموعه مترادف از آن دو اسم ورودی گرفته شده است. آزمایش‌های انجام‌شده نشان داد که محاسبه شباهت برچسب پرسش‌ها، کارآمدی بازیابی اطلاعات را به صورت معناداری بهبود می‌بخشد.

هستی‌شناسی[2]

در طول یک دهه گذشته، هستی‌شناسی‌ها با سرعت شتابناکی در جوامع نظام اطلاعاتی رواج پیداکرده است. به‌طورکلی، هستی‌شناسی به عنوان مفهوم‌سازی یک دامنه در نظر گرفته می‌شود (گروبه[3]، ۱۹۹۳). آن معمولاً مفاهیم را در یک سلسله مراتب درهم‌تنیده سازماندهی می‌کند و از طریق مجموعه‌ای غنی از روابط معنایی آن‌ها را به هم پیوند می‌دهد. در هستی‌شناسی، یک اصطلاح به عنوان واحد معنی‌دار تعریف می‌شود که از محتوای یک یا چند واژه تشکیل‌شده است. هر اصطلاح در متون، دارای ویژگی‌های مشخصی است و ممکن است با انواع خاصی از روابط معنایی با گروهی از اصطلاحات دیگر مرتبط باشد. به طور ویژه، در بازیابی اطلاعات می‌توان اصطلاح‌های مرتبط به یک اصطلاح را با ردیابی روابط آن‌ها در رابطه با آن اصطلاح موردنظر به دست آورد. مک گرینس[4] (۱۹۹۸) با استفاده از یک نظام جستجوی دانش‌بنیان به نام جستجوگر[5] یک آزمایش را انجام داد. وی گزارش داد که در صورت وجود شرایط خاص، هستی‌شناسی‌ها عملکرد جستجوی ابزارها را ازنظر بازیافت، دقت و سهولت شکل‌گیری پرس‌وجو بهبود می‌بخشند.

ساختار رده‌بندی[6]

لوئیس و جونز (۱۹۹۶) ایده ساختار «طبقه‌بندی کننده» را بر اساس یک کل ارائه دادند تا روابط پارادایمی بین اصطلاحات را نشان دهند و به جایگزینی واژه‌های کنترل‌شده در نمایه‌سازی و جستجوی مبتنی بر پردازش زبان طبیعی اجازه دهند. به عنوان مثال، نمادهای موجود در پایگاه‌های دانش، پایگاه‌های نظام خبره[7]، دیکشنری داده‌ها و کد منبع معمولاً نام‌هایی بگیرند که کلمات یا ترکیبات زبان طبیعی هستند. روابط بین اصطلاحات موجود در این ساختارها ممکن است مفیدتر از یک اصطلاح‌نامه عمومی برای بازیابی متون در یک دامنه خاص باشد.

1. Smeaton
2. Ontology
3. Gruber
4. McGuinness
5. FINDUR
6. Classificatory” structure
7. Expert-system

رویکرد استخراجی

از دید کاربر، اطلاعات مربوطه موجود در مدارک متنی- نه خود کلمات کلیدی یا مدارک- هدف نهایی بازیابی اطلاعات است. یک نظام بازیابی اطلاعات باید مبتنی بر مفهوم باشد نه مبتنی بر مدارک. بنابراین، موضوع استخراج اطلاعات مطرح می‌شود.

استخراج واقعیت و مدخل

استخراج اطلاعات سطح پایین مانند استخراج موجودیت نامدار، اغلب یک جزء ضروری در کنترل بیشتر انواع پرسش‌های زبان طبیعی است. از سال ۱۹۹۹، مسیر پاسخ دادن به پرسش‌ها در اجلاس بازیابی متن (ترک) بر ارزیابی عملکرد نظام‌های پرسش و پاسخ متمرکزشده است. یک پرسش زبان طبیعی مانند «چه زمانی نیکسون از چین دیدار کرد؟» نظام‌های شرکت‌کننده در اجلاس ملزم به بازیابی پاسخ‌های صحیح (جمله یا پاراگراف کوتاه به‌جای کل مدارک) از مجموعه مدارک بودند. استخراج موجودیت نامدار برای مشخص کردن پاسخ به پرسش‌های چه کسی و چه زمانی مفید است. برای پاسخگویی به پرسش‌های کجا، به یک پایگاه داده بزرگ جغرافیایی نیاز است.

استخراج قالب پرسش[1]

در رویکرد گسترشی در اجلاس بازیابی متن، پرسش‌های مربوط به زبان طبیعی و جملات موجود در مدارک در ساختارهای از پیش تعیین‌شده الگوهای مشابه تجزیه می‌شوند. شکاف‌های موجود در قالب‌ها شامل نوع پاسخ، هسته پرسش و موجودیت‌های نامدار و غیره است. سپس نظام، الگوی پرسش ثبت‌شده را با الگوی جمله موجود در مدرک مطابقت داده و جمله یا پاراگرافی که بر اساس شباهت‌های قالب آن‌ها احتمال بیشتری دارد که پاسخ پرسش مورد نظر باشد، بازیابی می‌شود. این رویکرد مبتنی بر قالب، قوانین اکتشافی را با استخراج اطلاعات برای تولید قالب‌های محتوا ادغام می‌کند که اندازه‌گیری شباهت را بر اساس برخی از سرنخ‌های زبانی امکان‌پذیر می‌کند.

آسک‌جوز[2] فناوری پردازش زبان طبیعی با قضاوت ویراستاری انسانی را ادغام می‌کند تا یک خدمت پرسش-پاسخ را به جستجوگرهای اطلاعاتی آنلاین ارائه دهد. کاربران یک پرسش را به زبان انگلیسی ارائه می‌کنند. این پرسش از لحاظ معنای واژه، متن و دستور زبان توسط موتور پردازش پرسش تفسیر می‌شود. سپس، پرسش تفسیر شده در برابر مجموعه‌ای از الگوهای پرسش از پیش تعریف‌شده در یک پایگاه دانش مقایسه می‌شود. کاربر، سیاهه کوتاهی از پرسش‌های خاص را دریافت می‌کند و از وی خواسته می‌شود تا پرسشی را انتخاب کند که نیازهای اطلاعاتی وی را به بهترین شکل ممکن برآورده می‌کند. سرانجام ، نظام یادشده بر اساس انتخاب کاربر ، اطلاعات مرتبط را از پایگاه دانش

1. Question template extraction
2. AskJeeves

بازیابی می‌کند. با وجود این می‌توان استنباط کرد که آسک‌جوز واقعاً به پرسش‌هایی همانند نظام‌های موجود در اجلاس بازیابی متن پاسخ نمی‌دهد؛ زیرا آسک‌جوز به این امر نیاز دارد که کاربران از میان چندین پرسش پیشنهاد شده، گزینه‌هایی را انتخاب کنند که پرسش آن‌ها از قبل طراحی شده است.

رویکرد تحولی[1]

یک محصول عمده تجزیه و تحلیل مدارک در بازیابی اطلاعات نمایه است که برای دستیابی به بازیابی سریع و دقیق بسیار مهم است. نمایه‌ساز وظیفه اختصاص برچسب به مدارک متنی را برای اطمینان از شناسایی و بازیابی موارد مناسب در زمان مناسب بر عهده دارد. نمایه یک مدرک، شامل سیاهه‌ای از اصطلاحات انتخاب شده و اطلاعات مرتبط با آن‌ها است. یکی از گزینه‌های دیگر برای بهبود نمایه‌سازی مدارک، ایجاد نمایه‌هایی بر اساس دانش زبانی عمیق‌تر از مدارک است و نه تعداد دفعات فراوانی کلمات (سالتون، ۱۹۸۹). این واضح است که فنون پردازش زبان طبیعی می‌توانند مدارک را به بازنمون‌های بهتر برای اهداف نمایه‌سازی و جستجو در مقایسه با روش‌های ساده مبتنی بر کلمه تبدیل کنند (پرز کابولوی و استرزال کاوسکی[2]، ۱۹۹۹) که در ادامه به برخی از این فنون اشاره شده است.

عبارت سازی نحوی[3]

یکی از یافته‌های اجلاس بازیابی متن این بود که در حالت کلی عبارت‌سازی مفید است. به‌طور خاص، عبارت‌سازی نحوی یک پیشرفت احتمالی نسبت به عبارت‌سازی آماری است. در میان انواع مختلف عبارت‌سازی اسامی نحوی شامل عبارت‌های اسامی کامل، خرده عبارت‌های مجاور در عبارت‌های اسمی، جفت‌های هد- مادیفایر[4]، کلمات منفرد به علاوه هدمادیفایرها[5] فقط باعث مؤثرترین اجرا در «کلاریتیچ ان ال پی[6]» شدند. میانگین دقت حدود ۱۳ درصد نسبت به خود کلمات پایه افزایش‌یافته است (وهیز[7]، ۱۹۹۹). از آزمایش‌ها اصطلاح‌های نمایه‌سازی با انگیزه زبانی دریافت شده است که که افزایش مجموعه‌های نمایه با اصطلاحات مرکب منجر به بهبود معنی‌داری در اثربخشی بازیابی مدارک برای هر دو جفت مجاور و جفت اصلاح‌کننده سر می‌شود (آرامپاتزیس، وایده، کوستر ، و بومل[8]، ۲۰۰۰). بنابراین در نظام پرسش و پاسخ، بردارهای جداگانه برای عبارات فعلی و اسمی در رابطه با هر جمله در مدارک ایجاد می‌شود و بر اساس مجموعه ویژگی استخراج شده از پرسش نمره اختصاص می‌یابد (آیشمن و سرینیواسان[9] ، ۱۹۹۹).

1. Transformation Approach
2. Perez-Carballo & Strzalkowski
3. Syntactic phrasing
4. Head-modifier pairs
5. Headmodifier pairs
6. CLARITECH NLP
7. Voorhees
8. Arampatzis, Weide, Koster, & Bommel
9. Eichmann & Srinivasan

تحلیل وابستگی[۱]

برخی از فنون تجزیه و تحلیل وابستگی در یافتن پاسخ‌های صحیح به پرسش‌های کاربران در بازیابی اطلاعات کارآمد هستند. این‌ها شامل تجزیه وت حلیل نحوی کم‌عمق است که عبارات اصلی و وابستگی بین آن‌ها ، مانند موضوع–فعل–اشیاء و وضوح هسته[۲] را برای تاریخ‌ها و نام‌ها و موارد دیگر مشخص می‌کند (استرزال کاوسکی و دیگران، ۲۰۰۰). به عنوان مثال، می‌توان «چهارشنبه» یا «ژوئن» را به تاریخ خاص با وضوح هسته ترجمه کرد. همچنین، روابط وابستگی گرامری نیز اتخاذ شده است تا به‌طور خودکار کلمات مشابه بازیابی شود (لین[۳]، ۱۹۹۸).

تحلیل معنایی[۴]

نمایه‌های مدارک می‌تواند بر مبنای معنا و مفهوم باشد (کولانتس[۵]، ۱۹۹۴؛ هولس[۶]، ۱۹۹۴). به طور معمول، جملات موجود در مدارک ابتدا به ساختار درختی تجزیه می‌شوند. سپس برخی از قوانین گرامری و واژه‌نامه‌ها برای تشخیص عبارت‌ها به عنوان موجودیت‌ها به کار گرفته می‌شود. سپس برخی از روابط معنایی بین موجودیت‌ها شناسایی‌شده و در پایگاه داده ذخیره می‌شوند. در نهایت روابط معنایی در پرسش مطرح شده با رکوردهای پایگاه داده مطابقت داده می‌شود (کتز[۷]، ۱۹۹۷؛ لیت کوسکی[۸]، ۱۹۹۹). پروژه پژوهش‌های «فریم نت[۹]» در حال ساختن یک منبع واژگانی است که هدف آن ارائه جمله‌های حاشیه‌نویسی‌شده معنایی و نحوی است که از آن می‌توان اطلاعات موثق در مورد ظرفیت‌ها یا احتمالات ترکیبی موارد در نظر گرفته‌شده برای تجزیه و تحلیل را گزارش کرد (فیل مول و بیکر[۱۰]، ۲۰۰۱). انتظار می‌رود که داده‌های «فریم نت» به ابهام‌زدایی حس کلمات، ترکیب معنایی، انتخاب‌های معتبر از بین بخش‌های رقیب و فعال کردن واژگان مرتبط با موضوع به منظور شناخت و انتخاب حس در قسمت‌های متوالی یک متن از طریق عضویت در قاب مشترک کمک کنند.

فرمالیسم منطقی[۱۱]

متون از طریق کارکردهای خاص که نوعی از قوانین تولیدی هستند به فرمول‌های منطقی ترجمه می‌شوند. این کارکردها به ساختار نحوی جملات بستگی دارند (سوود و بولک[۱۲]، ۱۹۷۸). بازیابی را می‌توان فرایندی برای یافتن پیوندهای متغیر

1. Dependency analysis
2. Coreference resolution
3. Lin
4. Semantic analysis
5. Collantes
6. Hull
7. Katz
8. Litkowski
9. FrameNet
10. Fillmore & Baker
11. Logic formalism
12. Schwind & Bolc

دانست که نمایش منطقی یک پرسش به زبان طبیعی را برآورده می‌کند. عیب این روش این است که استنتاج پیچیده است. امروزه رویکرد اتحادی چیزی بیش از آرمان‌گرایی است که بسیار مطلوب‌تر از واقع‌گرایی است؛ اگرچه تلاش‌هایی برای رسیدن به این هدف انجام شده است. پایگاه دانش حس مشترک سیس[1] یکی از محصولات پژوهشی پیشگام است که هدف آن ایجاد مبنایی برای دانش عقل سلیم است (لنت[2]، ۱۹۹۵). فناوری‌های پردازش زبان طبیعی قابل استفاده در چهارچوب «اِن اِل پی آی آر» محدود به مثال‌های توصیف‌شده نیستند بلکه، این مثال‌ها ارائه شده‌اند تا نشان دهند چهارچوب ارائه شده می‌تواند پژوهش‌های گسترده‌ای را در استفاده از پردازش زبان طبیعی در بازیابی اطلاعات پشتیبانی کند. لازم به ذکر است که فنون انفرادی همیشه در نظام‌های واقعی بازیابی اطلاعات به هم پیوسته‌اند. برای مثال «ان ال بی» یک نظام چندرسانه‌ای برای یادگیری الکترونیکی است (زاهانگ و نونامیکر[3]، ۲۰۰۰) که به کاربران امکان می‌دهد پرسش خود را به طبیعی ارائه کنند و سپس کلیپ‌های ویدیویی مناسب را که احتمالاً شامل پاسخ پرسش‌هاست، بازیابی کنند. این نظام، تجزیه و تحلیل تکواژ شناختی، استخراج موجودیت نامدار، گسترش مبتنی بر اصطلاح‌نامه و بازنمایی معنایی مبتنی بر فریم را ادغام می‌کند.

بحث و بررسی

محدودیت‌های جستجو مبتنی بر کلمه کلیدی سنتی بسیار زیاد است؛ زیرا کیفیت اطلاعات مهمتر از کمیت آن است. پردازش زبان طبیعی رویکردی است که اغلب برای رفع مشکلات بازیابی اطلاعات سنتی و بهبود عملکرد نظام‌های بازیابی اطلاعات استفاده می‌شود. با گسترش سریع فناوری شبکه در سال‌های اخیر، بازیابی اطلاعات، روند نمایه‌سازی متن و جستجوی مدارک مفید را در یک محیط توزیع شده نشان داده است. این امر باعث افزایش اهمیت، توسعه و ایجاد چهارچوبی می‌شود که این چهارچوب بتواند از پژوهش‌های بازیابی اطلاعات ناهمگن با ادغام پردازش زبان طبیعی با فناوری‌های بازیابی اطلاعات به صورت گسترده پشتیبانی کند. چهارچوب «اِن اِل پی آی آر» در راستای دستیابی به هدف یاد شده طراحی‌شده است. آن به صورت مناسبی پردازش زبان طبیعی را با بازیابی اطلاعات ادغام کرده است و این ادغام را در پنج رویکرد عمومیت بخشیده است. با ارائه مثال‌های دقیق از پردازش زبان طبیعی در چهارچوب یاد شده، ما نشان داده‌ایم که چهارچوب یاد شده می‌تواند از پژوهش‌ها و توسعه بازیابی اطلاعات در حال انجام پشتیبانی کند. افزون بر این، اهمیت چهارچوب «اِن اِل پی آی آر» می‌تواند فراتر از کارهای موجود باشد و بینشی در مورد جهت‌های بالقوه پژوهش‌های آینده و کاربردهای عملی در بازیابی اطلاعات ارائه دهد. ما در این مقاله درباره تلفیق پردازش زبان طبیعی در بازیابی اطلاعات بحث کرده‌ایم. به هر حال، برخی از پژوهشگران درباره استفاده از پردازش زبان طبیعی به جای تحلیل‌های آماری به عنوان ارزش‌افزوده برای بازیابی اطلاعات یاد کرده‌اند (اسپارک جونز، ۱۹۹۹). برخی از پژوهشگران

1. Cyc Common Sense Knowledge Base
2. Lenat
3. Zhang & Nunamaker

دریافتند که فنون پردازش زبان طبیعی فقط در برخی موارد به بهبود عملکرد بازیابی اطلاعات در سطح گسترده کمک می‌کند (ساندرسون، ۱۹۹۴). ما معتقدیم که رویکردهای پردازش زبان طبیعی و آماری باید مطابق سناریوهای مختلف انتخاب شوند. پردازش زبان طبیعی نه تنها از پژوهش‌های بازیابی اطلاعات بهره می‌برد بلکه، در عمل نیز آن را به کار می‌گیرد. با ادامه دادن به دستاوردهای پردازش زبان طبیعی و ادغام مناسب رویکردهای مختلف، همچنان چه در چهارچوب «اِن اِل پی آی آر» طراحی‌شده است، ما فکر می‌کنیم که پردازش زبان طبیعی نقش خیلی مهمی در آینده بازیابی اطلاعات ایفا خواهد کرد.

Reference

Abberley, D., Renals, S., & Cook, G. (1998). Retrieval of broadcast news documents with the THISL system. Proceedings of the TREC-7, Gaithersburg, MD.

Arampatzis, A., Weide, T., Koster, C., & Bommel, P. (2000). Linguistically motivated information retrieval. Encyclopedia of library and information Science (Vol. 69, pp. 201–222). New York: Marcel Dekker, Inc.

Belew, R. K. (1989). Adaptive information retrieval. Proceedings of the 12th annual international ACM/SIGIR conference on research and development in information retrieval (pp. 11–20). Cambridge, MA.

Bookstein, A., & Swanson, D. R. (1975). Probabilistic models for autonomic indexing. Journal of the American Society for Information Science, 26, 45–50.

Chen, H. (1995). Machine learning for information retrieval: Neutral networks, symbolic learning, and genetic algorithms. Journal of the American Society for Information Science, 46, 194 –216.

Chen, H., & Dhar, V. (1991). Cognitive process as a basis for intelligent retrieval systems design. Information Processing and Management, 27.۴۳۲ –۴۰۵,

Collantes, L.Y. (1994). Degree of agreement in naming objects and conepts for information retrieval. Journal of the American Society for Information Science, 46, 116 –132.

Eichmann, D., & Srinivasan, P. F. (1999). Filters, webs and answers: The University of Iowa TREC-8 results. In E.M. Voorhees & D.K. Harman (Eds.), Proceedings of the TREC-8 (pp. 259 –266). Gaithersburg, MD.

Fillmore, C.J., & Baker, C.F. (2001). Frame semantics for text understanding. Proceedings of the WordNet and Other Lexical Resources Workshop, NAACL, Pittsburgh, PA.

Gruber, T. R. (1993). Toward principles for the design of ontologies used for knowledge sharing (Knowledge Systems Laboratory, Technica Report KSL-93-04). Stanford, CA: Stanford Unviersity.

Hawking, D., Craswell, N. C., Thistlewaite, P., & Harman, D. (1999). Results and challenges in Web search evaluation. Proceedings of the Toronto ’99 (pp. 243–252). Canada.

Hui, B. (1998). Applying NLP to IR: Why and how [On-line]. Available: http: //citeseer.nj.nec.com/135389.html.

Hull, D. (1994). Improving text retrieval for the routing problem using latent semantic indexing. Proceedings of the 17th annual international ACM/SIGIR conference on research and development in information retrieval (pp. 282–292). Dublin.

Jennings, A., & Higuichi, H. (1992). A browser with a neural network user model. Library Hi Tech, 10, 77–93.

Katz, B. (1997). From sentence processing to information access on the World Wide Web. Proceedings of the AAAI Spring Symposium on Natural Language Processing for the World Wide Web (pp. 77– 86).Stanford, CA.

Korfhage, R. R. (1997). Information storage and retrieval. New York: John Wiley & Sons, Inc.

Lenat, D. B. (1995). CYC: Toward programs with common sense. Communications of ACM, 33, 30 – 49.

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

Lin, D. (1998). Automatic retrieval and clustering of similar words. Proceedings of the COLING-ACL ’98 (pp. 768 –774). Montreal, Canada.

Litkowski, K. C. (1999). Question-answering using semantic relation triples. Proceedings of the TREC-8 (pp. 267–274). Gaithersburg, MD.

Maron, M. E., & Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. Journal of the ACM, 7, 216 –243.

McGuinness, D. (1998). Ontological issues for knowledge-enhanced search. In N. Guarino (Ed.), Proceedings of the formal ontology in information systems (pp. 302–316). Trento, Italy: IOS Press.

Miller, D., Leek, T., & Schwartz, R. M. (1998). BBN at TREC7: Using hidden markov models for information retrieval. Proceedings of the TREC-7 (pp. 133–142). Gaithersburg, MD.

Miller, G. A. (1990). WORDNET: An on-line lexical database. International Journal of Lexicography, 3– 4, 235–312.

Paice, C. D. (1990). Another stemmer. SIGIR Forum, 24, 56 – 61.

Perez-Carballo, J., & Strzalkowski, T. (1999). Natural language information retrieval: Progress report. Information processing and Management,36, 155–178.

Porter, M. F. (1980). An algorithm ofr suffix stripping. Program, 14(3),130 –137.

Rau, L. F., & Jacobs, P. S. (1989). ir: Natural language for information retrieval. International Journal of Intelligent Sytems, 4, 319 –343.

Salton, G. (1989). Automatic text processing. Reading, MA: AddisonWesley Publishing Company, Inc.

Sanderson, M. (1994). Word sense disambiguation and information retrieval. Proceedings of the 17th ACM International Conference on Research and Development in Information Retrieval (pp. 142–151).Dublin, Ireland.

Schwind, C., & Bolc, L. (1978). A formalism for the description of question answering systems. In L. Bolc (Ed.), Natural language communication with computers (63, pp. 1– 47). New York: Springer.

Smeaton, A. F. (1999). Using NLP or NLP resources for information retrieval tasks. In T. Strzalkowski (Ed.), Natural language information retrieval (pp. 99 –111). Dordrect, The Netherlands: Kluwer Academic Publishers.

Sparck-Jones, K. (1999). What is the role for NLP in text retrieval. In T. Strzalkowski (Ed.), Natural language information retrieval (pp. 1–25). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Srihari, R., & Li, W. (1999). Information extraction supported question answering. Proceedings of the TREC-8 (pp. 185–196). Gaithersburg, MD.

Strzalkowski, T., Perez-Carballo, J., Karlgren, J., Hulth, A., Tapanainen, P., & Lahtinen, T. (1999). Natural language information retrieval: TREC-8 report. Proceedings of the TREC-8 (pp. 381–390). Gaithersburg, MD.

Strzalkowski, T., Stein, G.C., Wise, B., & Bagga, A. (2000). Towards the next generation information retrieval. Proceedings of the 6th Conference on content-based multimedia information access (RAIO'00) (pp. 1196 –1207). Paris, France.