تکنیک‌ها، کاربردها و پیشرفت‌های استخراج اطلاعات از متن

محمدرضا بحرانی
دانشجو کارشناسی ارشد، گروه مهندسی کامپیوتر، موسسه آموزش عالی آپادانا، شیراز
هاله همایونی
استادیار، عضو هیئت علمی گروه مهندسی کامپیوتر، موسسه آموزش عالی آپادانا، شیراز
زهرا تصمیم قطعی
کارشناس ارشد گروه مهندسی کامپیوتر، موسسه آموزش عالی آپادانا، شیراز

چکیده

استخراج اطلاعات (Information Extraction) یکی از حوزه‌های کلیدی پردازش زبان طبیعی است که هدف آن استخراج داده‌های ساختاریافته از متون غیرساختاریافته است. این مقاله مروری به بررسی پیشرفت‌های اخیر در تکنیک‌های مختلف استخراج اطلاعات، از جمله شناسایی موجودیت های نامدار (NER)، استخراج روابط، شناسایی رویدادها و تحلیل احساسات می‌پردازد. همچنین، چالش‌های مرتبط با داده‌های حجیم و متنوع، استفاده از مدل‌های یادگیری عمیق و تکنیک‌های نوظهور مانند مدل‌های زبانی بزرگ (LLMs) بررسی شده است. در نهایت، کاربردهای گسترده این حوزه در سیستم‌های دانش‌محور، شبکه‌های اجتماعی و ابزارهای تحلیل داده مورد توجه قرار گرفته است.

واژگان کلیدی: استخراج اطلاعات، شناسایی رویدادها، استخراج روابط، تحلیل احساسات، مدل‌های زبانی بزرگ.

مقدمه

مقالات استخراج اطلاعات یکی از شاخه‌های مهم و رو به رشد در حوزه پردازش زبان طبیعی و علوم داده است که هدف آن، استخراج داده‌های ساختاریافته و مفید از متون غیرساختاریافته است. با افزایش روزافزون حجم داده‌های متنی در منابع دیجیتال نظیر شبکه‌های اجتماعی، مقالات علمی، پایگاه‌های خبری و سایر منابع، نیاز به ابزارهایی که بتوانند این داده‌ها را تحلیل و ساختارمند کنند، بیش از پیش احساس می‌شود. در این راستا، استخراج اطلاعات به عنوان یک فناوری کلیدی، نقش مهمی در پردازش و استفاده بهینه از این حجم عظیم داده ایفا می‌کند.

یکی از اصلی‌ترین کاربردهای استخراج اطلاعات، شناسایی موجودیت‌های نامدار است که به تشخیص و دسته‌بندی موجودیت‌هایی نظیر اشخاص، مکان‌ها، سازمان‌ها و تاریخ‌ها در متن می‌پردازد. این تکنیک به صورت گسترده در زمینه‌هایی نظیر موتورهای جستجو، ترجمه ماشینی و سیستم‌های پرسش و پاسخ مورد استفاده قرار می‌گیرد. علاوه بر این، استخراج روابط نیز یکی دیگر از زیرحوزه‌های مهم استخراج اطلاعات است که به شناسایی روابط معنایی میان موجودیت‌ها در متن می‌پردازد و نقش کلیدی در ساخت و توسعه گراف‌های دانش ایفا می‌کند.

در سال‌های اخیر، پیشرفت‌های حاصل شده در زمینه یادگیری ماشین و به ویژه یادگیری عمیق تحول عظیمی در حوزه استخراج اطلاعات ایجاد کرده است. استفاده از مدل‌های زبانی پیشرفته مانند BERT و مدل‌های زبانی بزرگ (LLM) نظیر GPT، امکان تحلیل دقیق‌تر و انعطاف‌پذیری بیشتر را در پردازش زبان طبیعی فراهم کرده است. این مدل‌ها با توانایی خود در درک عمیق متون، نه تنها باعث بهبود عملکرد سیستم‌های استخراج اطلاعات شده‌اند، بلکه کاربردهای جدیدی مانند شناسایی احساسات و شناسایی رویدادها را نیز تسهیل کرده‌اند.

یکی از چالش‌های اصلی در این حوزه، مدیریت داده‌های حجیم و متنوع است. منابع متنی مختلف، از جمله مقالات علمی، شبکه‌های اجتماعی، و محتوای خبری، دارای تفاوت‌های اساسی در ساختار و زبان هستند. علاوه بر این، وجود داده‌های نویزی و غیرساختاریافته در این منابع، فرآیند استخراج اطلاعات را پیچیده می‌کند. برای مقابله با این چالش‌ها، استفاده از روش‌های ترکیبی که یادگیری نظارت‌شده و یادگیری بدون نظارت را ترکیب می‌کنند، به‌طور گسترده مورد توجه قرار گرفته است.

کاربردهای استخراج اطلاعات بسیار گسترده است و از تحلیل شبکه‌های اجتماعی گرفته تا سیستم‌های پیشنهاددهنده و تحلیل احساسات کاربران را شامل می‌شود. برای مثال، در حوزه پزشکی، این فناوری می‌تواند برای استخراج اطلاعات حیاتی از متون تحقیقاتی یا پرونده‌های پزشکی بیماران مورد استفاده قرار گیرد. در حوزه بازاریابی، تحلیل احساسات کاربران نسبت به محصولات یا خدمات می‌تواند به شرکت‌ها در بهبود استراتژی‌های خود کمک کند. همچنین، شناسایی و پیش‌بینی رویدادها در شبکه‌های اجتماعی، ابزار قدرتمندی برای مدیریت بحران‌ها فراهم می‌کند.

این مقاله مروری تلاش دارد تا با ارائه یک دیدگاه جامع نسبت به پیشرفت‌های اخیر در زمینه استخراج اطلاعات، روش‌ها، ابزارها و چالش‌های این حوزه را بررسی کند. همچنین، کاربردهای کلیدی این فناوری در حوزه‌های مختلف و مسیرهای آینده تحقیقاتی مورد بحث قرار خواهد گرفت. این مرور نه تنها به شناسایی نقاط قوت و ضعف رویکردهای موجود کمک می‌کند، بلکه زمینه‌ساز ارائه ایده‌های نوین برای تحقیقات آینده خواهد بود.

مبانی و پیشینه استخراج اطلاعات

استخراج اطلاعات به عنوان یکی از حوزه‌های بنیادی پردازش زبان طبیعی، با هدف تبدیل داده‌های غیرساختاریافته متنی به اطلاعات ساختاریافته مطرح شده است. این حوزه در دهه‌های گذشته، با تمرکز بر روش‌های مبتنی بر قواعد و استفاده از الگوهای استخراج، شکل گرفت. در این روش‌های اولیه، تحلیل‌های زبانی و طراحی دستی قواعد، ابزار اصلی برای شناسایی موجودیت‌ها و روابط در متن بود. هرچند این روش‌ها به دلیل نیاز به دخالت انسانی و محدودیت‌های آن در پردازش داده‌های متنوع، کارایی محدودی داشتند.
در دهه ۱۹۹۰، با ظهور یادگیری ماشین، تحولی اساسی در استخراج اطلاعات به وجود آمد. روش‌های آماری و الگوریتم‌های یادگیری نظارت‌شده امکان استفاده از داده‌های برچسب‌دار برای آموزش مدل‌های قوی‌تر را فراهم کردند. ابزارهایی مانند HMM و CRF در این دوره برای شناسایی موجودیت‌های نامدار و استخراج روابط مورد استفاده قرار گرفتند. با این حال، این مدل‌ها همچنان نیازمند طراحی ویژگی‌های دستی توسط متخصصان بودند.

در سال‌های اخیر، پیشرفت‌های چشمگیر در یادگیری عمیق و معرفی مدل‌های زبانی پیشرفته نظیر BERT و GPT، مرزهای جدیدی در حوزه استخراج اطلاعات گشودند. این مدل‌ها با توانایی ظویژگی‌های دستی را کاهش داده‌اند، بلکه دقت و قابلیت تعمیم‌پذیری مدل‌ها را به طور چشمگیری افزایش داده‌اند. به‌علاوه، تکنیک‌هایی نظیر یادگیری انتقالی و استفاده از داده‌های بزرگ برای پیش‌آموزش مدل‌ها، امکان پردازش زبان‌های کم‌منبع را نیز فراهم کرده است.

امروزه، استخراج اطلاعات به عنوان یکی از ستون‌های اصلی کاربردهای متنوع مانند ساخت گراف‌های دانش، تحلیل شبکه‌های اجتماعی، و سیستم‌های پرسش و پاسخ مطرح است. توسعه روش‌ها و ابزارهای جدید، نه تنها این حوزه را به سمت دقت بیشتر سوق داده، بلکه امکان پردازش داده‌های حجیم و غیرساختاریافته را نیز فراهم کرده است. این روند تحول‌آفرین نشان می‌دهد که استخراج اطلاعات، یکی از ابزارهای کلیدی در توسعه هوش مصنوعی و مدیریت داده در عصر دیجیتال باقی خواهد ماند.

مفاهیم اولیه در استخراج اطلاعات

استخراج اطلاعات، فرآیند شناسایی و استخراج داده‌های ساختاریافته از متون غیرساختاریافته اطلاق میشود. ای داده‌ها میتوانند شامل موجودیت‌ها، روابط، رویدادها و سایر اطلاعات مفید باشند که به شکل ساختاریافته برای کاربردهای مختلف پردازش می‌شوند. در این بخش، مفاهیم اصلی و پایه‌ای که در استخراج اطلاعات مورد استفاده قرار می‌گیرند، معرفی می‌شوند.

- موجودیت‌های نامدار(Named Entities)

موجودیت‌های نامدار شامل عناصر خاصی از متن هستند که به اشخاص، مکان‌ها، سازمان‌ها، تاریخ‌ها، اعداد و سایر موارد مشابه اشاره دارند. شناسایی موجودیت‌های نامدار اولین گام مهم در فرآیند استخراج اطلاعات است و هدف آن برچسب‌گذاری این موجودیت‌ها در متن و دسته‌بندی آن‌ها به انواع مشخص است.

- روابط

روابط نشان‌دهنده پیوندهای معنایی میان موجودیت‌ها هستند. برای مثال، در جمله "البرت انیشتین در آلمان متولد شد"، رابطه "محل تولد" میان دو موجودیت "البرت انیشتین" و "آلمان" برقرار است. استخراج روابط یکی از وظایف کلیدی در ایجاد ساختار معنایی از متون است.

- رویدادها

رویدادها نشان‌دهنده وقوع یک عمل یا پدیده خاص در یک بازه زمانی مشخص هستند. شناسایی رویدادها شامل استخراج اطلاعاتی نظیر فاعل، مفعول، زمان و مکان رویداد می‌باشد. به عنوان مثال، در جمله "زلزله‌ای در تهران رخ داد"، رویداد "زلزله" و مکان "تهران" مورد شناسایی قرار می‌گیرند.

- تحلیل احساسات

رویدادها تحلیل احساسات به شناسایی و دسته‌بندی احساسات یا نگرش‌های بیان شده در متن می‌پردازد. این مفهوم به طور خاص برای درک نگرش کاربران نسبت به محصولات، خدمات یا رویدادها به کار می‌رود.

- استخراج اطلاعات باز

این روش به استخراج تمامی روابط ممکن در متن، بدون نیاز به تعریف پیشینی روابط خاص، می‌پردازد. استخراج اطلاعات باز برای کاربردهایی که نیاز به تحلیل گسترده‌تر و بدون محدودیت دارند، بسیار مفید است.

- داده‌های غیرساختار یافته

این داده‌ها شامل متونی است که قالب و ساختار مشخصی ندارند، مانند مقالات، توییت‌ها یا ایمیل‌ها در این دسته قرار دارند.

- داده‌های ساختار یافته

داده‌هایی هستند که در قالب مشخصی نظیر جداول پایگاه داده سازماندهی شده‌اند. استخراج اطلاعات تلاش می‌کند داده‌های غیرساختاریافته را به داده‌های ساختاریافته تبدیل کند.

– داده‌های ارزیابی دقت و عملکرد

داده‌هایی برای ارزیابی کارایی سیستم‌های استخراج اطلاعات، معیارهایی نظیر دقت، فراخوانی و امتیاز استفاده می‌شوند. این معیارها نشان‌دهنده توانایی سیستم در شناسایی صحیح و جامع اطلاعات هدف هستند.

– کاربرد ابزارهای زبانی

استخراج اطلاعات معمولا نیازمند استفاده از ابزارهایی نظیر تجزیه کننده‌های نحوی، برچسب‌زن‌های نقش کلمات و مدل‌های معنایی برای درک بهتر متن است. این ابزارها اساس پردازش زبان طبیعی در سیستم‌های استخراج اطلاعات را تشکیل می‌دهند.

خلاصه روند پیشرفت در استخراج اطلاعات

استخراج اطلاعات به عنوان یک حوزه علمی از دهه ۱۹۷۰ میلادی ظهور کرد، زمانی که پژوهشگران تلاش می‌کردند تا داده‌های خاصی را از متون غیرساختاریافته استخراج کنند. در ابتدا، روش‌های مبتنی بر قواعد و الگوهای دستی، مانند سیستم‌های تطبیق الگو، به کار گرفته شدند. این روش‌ها به دلیل نیاز به طراحی دستی و محدودیت در پوشش دامنه‌های مختلف، کارایی محدودی داشتند. در دهه ۱۹۹۰، ظهور روش‌های آماری و یادگیری ماشین نظیر مدل‌های مارکوف پنهانو ماشین‌های بردار پشتیبان، پیشرفت قابل توجهی را در این حوزه به همراه داشت. این مدل‌ها امکان استفاده از داده‌های برچسب‌دار را برای یادگیری خودکار فراهم کردند و دقت استخراج اطلاعات را بهبود بخشیدند. در دهه‌های بعد، یادگیری عمیق و استفاده از شبکه‌های عصبی پیشرفته مانند LSTM و ترانسفورمرها نظیر BERT و GPT، انقلابی در این حوزه ایجاد کردند. این مدل‌ها توانایی پردازش داده‌های پیچیده و متنوع را با دقت و سرعت بالا فراهم

کردند. امروزه، با استفاده از مدل‌های زبانی بزرگ و یادگیری انتقالی، استخراج اطلاعات به ابزاری قوی برای تحلیل داده‌های حجیم و غیرساختاریافته تبدیل شده و کاربردهای گسترده‌ای در حوزه‌هایی مانند پزشکی، بازاریابی، و تحلیل شبکه‌های اجتماعی یافته است.

مروری بر روش‌های سنتی و مدرن در استخراج اطلاعات

- روش‌های سنتی در استخراج اطلاعات

روابط روش‌های سنتی در استخراج اطلاعات عمدتاً بر مبنای قواعد دستی و الگوهای از پیش تعریف‌شده طراحی شده بودند. این روش‌ها با استفاده از تجزیه‌های نحوی، تحلیل زبان‌شناختی، و قواعد از پیش تعیین‌شده تلاش می‌کردند موجودیت‌ها و روابط را از متون استخراج کنند. به عنوان مثال، سیستم‌هایی مانند FASTUS و AutoSlog از تکنیک‌های مبتنی بر الگو برای پردازش متون استفاده می‌کردند. این سیستم‌ها، گرچه در دامنه‌های محدود و متون تخصصی عملکرد خوبی داشتند، اما به دلیل نیاز به طراحی دستی الگوها و عدم انعطاف‌پذیری در مواجهه با داده‌های متنوع، محدودیت‌های قابل توجهی داشتند. علاوه بر این، روش‌های آماری مانند مدل‌های مارکوف پنهان (HMM) و ماشین‌های بردار پشتیبان (SVM) در اواخر دهه ۱۹۹۰ معرفی شدند که با استفاده از داده‌های برچسب‌دار و ویژگی‌های از پیش تعریف‌شده توانستند برخی از این محدودیت‌ها را کاهش دهند.

- ظهور روش‌های مدرن در استخراج اطلاعات

با پیشرفت یادگیری عمیق و افزایش توان پردازشی سیستم‌های کامپیوتری، روش‌های مدرن جایگزین رویکردهای سنتی شدند. روش‌های مدرن عمدتاً بر پایه شبکه‌های عصبی عمیق و مدل‌های زبانی قدرتمند طراحی شده‌اند. یکی از نخستین مدل‌های مدرن در این حوزه، استفاده از شبکه‌های عصبی بازگشتی (RNN) و نسخه‌های پیشرفته‌تر آن نظیر LSTM و GRU بود که برای پردازش داده‌های متوالی مانند متون بسیار مناسب بودند. این مدل‌ها به استخراج اطلاعات از داده‌های پیچیده‌تر، مانند متون غیرساختاریافته و نویزی، کمک کردند.

- نقش ترانسفورمرها و مدل‌های زبانی بزرگ

تحول بزرگ در روش‌های مدرن با معرفی معماری ترانسفورمر و مدل‌هایی مانند BERT و GPT آغاز شد. این مدل‌ها، با استفاده از مکانیزم توجه، توانایی تحلیل عمیق‌تری از متن را نسبت به مدل‌های قبلی فراهم کردند. برای مثال، مدل BERT با پیش‌آموزش بر روی حجم عظیمی از داده‌های متنی و استفاده از یادگیری انتقالی، امکان استخراج اطلاعات را با دقت بالاتر و بدون نیاز به برچسب‌گذاری گسترده فراهم کرد. مدل‌های زبانی بزرگ مانند GPT نیز، علاوه بر استخراج اطلاعات، توانایی تولید متن و تفسیر معنایی پیچیده‌تر را به این حوزه افزودند.

- ترکیب روش‌های سنتی و مدرن

اگرچه روش‌های مدرن در بسیاری از جنبه‌ها از روش‌های سنتی پیشی گرفته‌اند، اما ترکیب این دو رویکرد در برخی از حوزه‌ها موفقیت‌آمیز بوده است. برای مثال، در کاربردهایی که داده‌های کافی برای یادگیری عمیق در دسترس نیست، استفاده از قواعد سنتی به عنوان پیش‌پردازش یا تقویت مدل‌های مدرن می‌تواند عملکرد سیستم را بهبود بخشد. به‌علاوه، برخی از روش‌ها از ویژگی‌های استخراج‌شده توسط روش‌های سنتی به عنوان ورودی برای مدل‌های مدرن استفاده می‌کنند.

تکنیک‌ها و روش‌های استخراج اطلاعات

این تکنیک ها و روش‌های مختلفی برای استخراج اطلاعات به کار برده شده که هریک در نوع و روش خود موفق بوده‌اند. اما در این میان برخی از این روش‌ها به دلیل ساختار مناسب، برای اهداف خاص طراحی شده اند که به نوع خود استخراج اطلاعات را هدفمندتر کرده

است. در همین راستا، در متونی که ممکن است در رابطه با موضوع خاصی نوشته شده‌اند، این الگوریتم‌ها کارایی مناسب‌تری داشته باشند.

روش شناسایی موجودیت‌های نامدار

شناسایی موجودیت‌های نامدار (Named Entity Recognition-NER) یکی از وظایف اساسی در استخراج اطلاعات است که به شناسایی و دسته‌بندی موجودیت‌هایی مانند اشخاص، مکان‌ها، سازمان‌ها، تاریخ‌ها و موارد مشابه در متن می‌پردازد. NER از نقش مهمی در سیستم‌های پردازش زبان طبیعی نظیر ترجمه ماشینی، پرسش و پاسخ، و تحلیل متون برخوردار است. روش‌های مختلفی برای پیاده‌سازی NER وجود دارد که در این بخش به بررسی سه دسته اصلی می‌پردازیم.

- روش‌های مبتنی بر قواعد

در روش‌های مبتنی بر قواعد، شناسایی موجودیت‌ها بر اساس الگوها و قواعد زبان‌شناختی از پیش تعریف‌شده انجام می‌شود. این روش‌ها از ابزارهایی مانند تجزیه‌کننده‌های نحوی، الگوهای منظم، و واژگان‌های تخصصی برای شناسایی موجودیت‌ها استفاده می‌کنند. به عنوان مثال، یک قاعده ساده می‌تواند تشخیص موجودیت‌های عددی مانند تاریخ یا کدهای پستی باشد. این روش‌ها برای دامنه‌های خاص و متون با ساختار مشخص مناسب هستند، اما در مواجهه با داده‌های نویزی یا متنوع کارایی پایینی دارند. یکی از چالش‌های اصلی این روش، نیاز به طراحی و به‌روزرسانی دستی قواعد است که فرآیندی زمان‌بر و پرهزینه محسوب می‌شود.

- روش‌های مبتنی بر یادگیری ماشین

ظهور یادگیری ماشین انقلابی در حوزه NER ایجاد کرد. در این روش‌ها، مدل‌ها با استفاده از داده‌های برچسب‌دار آموزش می‌بینند و قادر به شناسایی الگوهای پیچیده‌تر در متن هستند. الگوریتم‌هایی نظیر مدل‌های مارکوف پنهان (HMM)، ماشین‌های بردار پشتیبان (SVM)، و جنگل‌های تصادفی از جمله روش‌های محبوب در این دسته به شمار می‌روند. این مدل‌ها ویژگی‌هایی نظیر نقش کلمات (POS Tags)، هم‌جواری کلمات، و فاصله‌های بین موجودیت‌ها را به عنوان ورودی دریافت می‌کنند. یکی از مزایای اصلی روش‌های یادگیری ماشین، کاهش وابستگی به قواعد دستی و توانایی تعمیم بهتر به داده‌های جدید است. با این حال، این روش‌ها همچنان نیازمند استخراج ویژگی‌های دستی بوده و ممکن است در پردازش زبان‌های کم‌منبع با چالش مواجه شوند.

- روش‌های استفاده از مدل‌های عمیق و ترانسفورمرها

در سال‌های اخیر، یادگیری عمیق و به ویژه استفاده از مدل‌های مبتنی بر ترانسفورمرها، استاندارد جدیدی در حوزه NER ایجاد کرده است. شبکه‌های عصبی بازگشتی (RNN) و نسخه‌های پیشرفته‌تر آن مانند LSTM و GRU، از اولین مدل‌های عمیقی بودند که برای NER استفاده شدند. این مدل‌ها با پردازش توالی کلمات، به شناسایی روابط پیچیده میان موجودیت‌ها پرداختند.

معرفی مدل‌های ترانسفورمر مانند BERT و RoBERTa، تحولی بزرگ در دقت و کارایی NER ایجاد کرد. این مدل‌ها با پیش‌آموزش بر روی مجموعه داده‌های بزرگ و استفاده از یادگیری انتقالی، توانایی درک بهتر متون و شناسایی موجودیت‌ها در زبان‌های مختلف را فراهم کردند. به عنوان مثال، BERT با استفاده از مکانیزم توجه قادر به درک وابستگی‌های طولانی‌مدت در متن است و عملکرد چشمگیری در شناسایی موجودیت‌های نامدار نشان داده است.

مدل‌های عمیق و ترانسفورمرها همچنین در شناسایی موجودیت‌های نامدار در زبان‌های کم‌منبع یا متون غیرساختاریافته موفق بوده‌اند. یکی دیگر از مزایای این مدل‌ها، قابلیت استفاده مجدد از مدل‌های پیش‌آموزش‌شده برای دامنه‌های مختلف با حداقل نیاز به داده‌های برچسب‌دار است.

روش استخراج روابط

استخراج روابط (Relation Extraction) یکی از وظایف کلیدی در استخراج اطلاعات است که هدف آن شناسایی و تعریف روابط معنایی میان موجودیت‌های شناسایی‌شده در متن است. این فرآیند درک عمیق‌تری از ارتباط میان عناصر مختلف متن را ممکن می‌سازد و نقش اساسی در ساختاردهی داده‌های غیرساختاریافته و تبدیل آن‌ها به داده‌های قابل پردازش ایفا می‌کند. در این بخش، دو رویکرد اصلی برای استخراج روابط و کاربردهای آن‌ها مورد بررسی قرار می‌گیرد.

- استخراج روابط از متون ساختاریافته و غیرساختاریافته

متون ساختاریافته، نظیر جداول پایگاه داده و اسناد فرم‌دار، معمولاً شامل اطلاعات از پیش سازماندهی‌شده‌ای هستند که استخراج روابط در آن‌ها آسان‌تر است. این روابط معمولاً به صورت مستقیم یا با استفاده از قوانین ساده قابل شناسایی هستند. به عنوان مثال، در پایگاه‌داده‌ای شامل اطلاعات افراد و شغل‌های آن‌ها، رابطه "شاغل بودن" میان افراد و موقعیت شغلی آن‌ها می‌تواند استخراج شود. در مقابل، متون غیرساختاریافته شامل داده‌هایی نظیر مقالات، ایمیل‌ها و پیام‌های شبکه‌های اجتماعی هستند که فاقد ساختار مشخصی هستند. در این حالت، استخراج روابط به ابزارهای پیشرفته پردازش زبان طبیعی نیاز دارد. الگوریتم‌های یادگیری ماشین و مدل‌های زبانی پیشرفته، مانند BERT و GPT، به‌طور گسترده برای استخراج روابط از این نوع داده‌ها استفاده می‌شوند. این مدل‌ها با استفاده از تحلیل معنایی و نحوی متن، روابط پیچیده‌تری را شناسایی می‌کنند. به عنوان مثال، در جمله "مارک زاکربرگ بنیان‌گذار فیسبوک است"، رابطه "بنیان‌گذار" میان "مارک زاکربرگ" و "فیسبوک" قابل استخراج است.با این حال، این روش‌ها همچنان نیازمند استخراج ویژگی‌های دستی بوده و ممکن است در پردازش زبان‌های کم‌منبع با چالش مواجه شوند.

- گراف‌های دانش و کاربردهای آن

گراف‌های دانش (Knowledge Graphs) یکی از ابزارهای کلیدی برای ذخیره و نمایش روابط استخراج‌شده هستند. در گراف‌های دانش، موجودیت‌ها به عنوان گره‌ها و روابط به عنوان یال‌های گراف نمایش داده می‌شوند. این ساختار، امکان نمایش اطلاعات به صورت بصری و همچنین تسهیل فرآیند جستجو و تحلیل داده‌ها را فراهم می‌کند.

گراف‌های دانش کاربردهای گسترده‌ای در حوزه‌های مختلف دارند. یکی از کاربردهای رایج آن‌ها، استفاده در موتورهای جستجو نظیر گوگل است که با کمک این گراف‌ها، پاسخ‌های دقیق‌تر و مرتبط‌تری به پرسش‌های کاربران ارائه می‌شود. در حوزه پزشکی، گراف‌های دانش می‌توانند برای نمایش روابط میان بیماری‌ها، علائم، و درمان‌ها به کار روند و در تحقیقات علمی و تصمیم‌گیری‌های درمانی مفید باشند. همچنین، در تحلیل شبکه‌های اجتماعی، این گراف‌ها به درک روابط میان کاربران و شناسایی تأثیرگذاران کلیدی کمک می‌کنند.

روش استخراج اطلاعات باز

استخراج اطلاعات باز (Open Information Extraction - Open IE) یکی از رویکردهای نوآورانه در حوزه استخراج اطلاعات است که به شناسایی و استخراج روابط و موجودیت‌ها از متون به صورت گسترده و بدون نیاز به تعریف از پیش تعیین‌شده روابط یا موجودیت‌ها می‌پردازد. این رویکرد برای پردازش داده‌های متنوع و غیرساختاریافته، مانند مقالات، اخبار و شبکه‌های اجتماعی، بسیار مناسب است و در مقایسه با استخراج اطلاعات هدفمند انعطاف‌پذیری بیشتری دارد.

- تفاوت‌های استخراج اطلاعات باز و هدفمند

تعریف روابط: در استخراج اطلاعات هدفمند، روابط و موجودیت‌ها از پیش تعریف شده و مدل برای شناسایی آن‌ها آموزش داده می‌شود. به عنوان مثال، رابطه "محل تولد" یا "بنیان‌گذار" به صورت دقیق مشخص می‌شود. در مقابل، استخراج اطلاعات باز به شناسایی تمامی روابط و موجودیت‌های ممکن در متن می‌پردازد و نیازی به تعریف از پیش ندارد.

انعطاف‌پذیری: در حالی که استخراج اطلاعات هدفمند تنها در دامنه‌های خاص با روابط از پیش تعیین‌شده کاربرد دارد، استخراج اطلاعات باز برای متون متنوع و بدون محدودیت دامنه بسیار مناسب است.

کاربرد: استخراج اطلاعات هدفمند معمولاً در سیستم‌هایی که نیاز به داده‌های دقیق و مشخص دارند، مانند سیستم‌های پزشکی یا مالی، استفاده می‌شود. اما استخراج اطلاعات باز برای کاربردهای گسترده‌تر، مانند ساخت گراف‌های دانش از متون عمومی، تحلیل شبکه‌های اجتماعی و موتورهای جستجو، کاربرد دارد.

- ابزارها و روش‌های جدید

در سال‌های اخیر، ابزارها و روش‌های جدیدی برای بهبود عملکرد استخراج اطلاعات باز توسعه یافته‌اند که برخی از مهم‌ترین آن‌ها عبارتند از:

۱. TEXTRUNNER: یکی از اولین سیستم‌های استخراج اطلاعات باز که بر اساس الگوهای ساده و روش‌های آماری عمل می‌کند. این سیستم با پردازش متون به صورت جمله به جمله، روابط و موجودیت‌ها را استخراج می‌کند.

۲. ReVerb: این ابزار با تمرکز بر استخراج روابط دقیق، از الگوهای نحوی ساده برای شناسایی افعال و عبارات وابسته به آن‌ها استفاده می‌کند. ReVerb در مقایسه با TEXTRUNNER دقت بالاتری در شناسایی روابط دارد.

۳. OLLIE: این ابزار با استفاده از یادگیری نظارت‌شده، توانایی استخراج روابط پیچیده و شرایط مرتبط با آن‌ها را دارد. OLLIE از مدل‌های زبانی برای درک بهتر متن بهره می‌برد.

۴. OpenIE 5.0: یکی از پیشرفته‌ترین ابزارهای استخراج اطلاعات باز که از معماری‌های مدرن ترانسفورمر برای پردازش زبان طبیعی استفاده می‌کند. این ابزار با بهره‌گیری از مدل‌هایی نظیر BERT و RoBERTa، دقت و انعطاف‌پذیری بیشتری را ارائه می‌دهد.

۵. Graphene: این ابزار با ترکیب استخراج اطلاعات باز و ساخت گراف‌های دانش، امکان نمایش بصری و تحلیل بهتر روابط استخراج‌شده را فراهم می‌کند.

روش شناسایی رویدادها

شناسایی رویدادها (Event Detection) یکی از حوزه‌های کلیدی در استخراج اطلاعات است که هدف آن شناسایی وقوع رویدادها، بازیابی جزئیات مرتبط و تحلیل داده‌های متنی برای درک بهتر روابط زمانی و مکانی رویدادها است. این فرآیند به‌ویژه در تحلیل داده‌های اجتماعی و خبری از اهمیت زیادی برخوردار است و کاربردهای گسترده‌ای در حوزه‌هایی مانند مدیریت بحران، تحلیل بازار و امنیت دارد.

- رویکردها در داده‌های اجتماعی و خبری

۱. رویکردهای مبتنی بر کلمات کلیدی و الگوها: در این روش‌ها، از کلمات کلیدی، عبارات خاص یا الگوهای نحوی برای شناسایی رویدادها در متن استفاده می‌شود. این رویکردها معمولاً مناسب برای داده‌های خبری با ساختار نسبتاً منظم هستند. برای مثال، در متنی خبری، عبارات نظیر "زلزله رخ داد" یا "اعتراض‌های گسترده" می‌توانند نشانگر وقوع رویداد باشند.

۲. مدل‌های آماری و یادگیری ماشین: این مدل‌ها با استفاده از داده‌های برچسب‌دار آموزش می‌بینند تا روابط میان واژگان و وقوع رویدادها را شناسایی کنند. الگوریتم‌هایی نظیر ماشین‌های بردار پشتیبان (SVM) و مدل‌های مارکوف پنهان (HMM) به طور گسترده در این زمینه مورد استفاده قرار گرفته‌اند.

۳. مدل‌های یادگیری عمیق و ترانسفورمرها: ظهور شبکه‌های عصبی عمیق و مدل‌های مبتنی بر ترانسفورمرها، نظیر BERT و RoBERTa، دقت و انعطاف‌پذیری در شناسایی رویدادها را به طور قابل توجهی افزایش داده است. این مدل‌ها قادر به شناسایی روابط پیچیده و استخراج اطلاعات از متون غیرساختاریافته و داده‌های اجتماعی هستند.

۴. تحلیل شبکه‌های اجتماعی: داده‌های اجتماعی نظیر توییت‌ها، پست‌های فیسبوک یا پیام‌های اینستاگرام اغلب شامل اطلاعات مرتبط با رویدادها هستند. تحلیل این داده‌ها نیازمند روش‌هایی است که با نویز بالا، اختصارات زبانی و تغییرات سریع داده سازگار باشند. مدل‌های پیشرفته یادگیری عمیق و ابزارهای NLP نظیر SpaCy و FastText در این زمینه کاربرد زیادی دارند.

روش تحلیل احساسات

در تحلیل احساسات (Sentiment Analysis) یکی از حوزه‌های مهم پردازش زبان طبیعی است که هدف آن شناسایی و ارزیابی نگرش، احساس یا عقاید بیان‌شده در متن‌ها است. این فرآیند به شناسایی قطبیت متن‌ها (مثبت، منفی یا خنثی) می‌پردازد و کاربردهای گسترده‌ای در تحلیل رفتار کاربران، بازاریابی و مدیریت ارتباط با مشتریان دارد. در این بخش، روش‌های اصلی تحلیل احساسات و کاربردهای آن در حوزه‌های مختلف بررسی می‌شود.

- روش‌های مبتنی بر واژگان و یادگیری ماشین

روش‌های مبتنی بر واژگان: روش‌های مبتنی بر واژگان از فرهنگ‌های لغوی استفاده می‌کنند که شامل لیستی از کلمات با قطبیت مشخص (مثبت یا منفی) هستند. در این روش، هر متن بر اساس حضور کلمات مثبت یا منفی و وزن آن‌ها تحلیل می‌شود. به عنوان مثال، واژه‌هایی نظیر "عالی" یا "وحشتناک" به ترتیب قطبیت مثبت و منفی دارند و در تحلیل احساسات متون تأثیرگذار هستند. این رویکرد به دلیل سادگی و نیاز کمتر به داده‌های آموزشی، در بسیاری از کاربردهای اولیه محبوب بود. با این حال، محدودیت‌هایی نظیر ناتوانی در شناسایی مفاهیم پیچیده، وجود دارد؛ برای مثال، جملات کنایه‌آمیز یا ترکیب کلمات که قطبیت واژگان را تغییر می‌دهند، به سختی تحلیل می‌شوند.

روش‌های مبتنی بر یادگیری ماشین: روش‌های مبتنی بر یادگیری ماشین، از داده‌های برچسب‌دار برای آموزش مدل‌هایی استفاده می‌کنند که توانایی شناسایی الگوهای پیچیده در متن را دارند. الگوریتم‌هایی نظیر ماشین‌های بردار پشتیبان (SVM)، بیز ساده و جنگل تصادفی (Random Forest) در این زمینه استفاده می‌شوند. این روش‌ها به دلیل قابلیت تعمیم‌پذیری و انعطاف‌پذیری، در بسیاری از سیستم‌های تحلیل احساسات به کار گرفته شده‌اند.

در سال‌های اخیر، استفاده از یادگیری عمیقو مدل‌های پیشرفته‌تر نظیر شبکه‌های عصبی بازگشتی (RNN)، شبکه‌های عصبی پیچشی (CNN) و مدل‌های مبتنی بر ترانسفورمرها (مانند BERT) تحول بزرگی در تحلیل احساسات ایجاد کرده است. این مدل‌ها با توانایی پردازش متون طولانی و درک روابط پیچیده میان کلمات، دقت بالاتری در شناسایی قطبیت متون ارائه می‌دهند.

- کاربردهای تحلیل احساسات در حوزه‌های مختلف
۱. بازاریابی و تحلیل برند: تحلیل احساسات نقش مهمی در درک نگرش کاربران نسبت به محصولات یا خدمات دارد. شرکت‌ها با تحلیل بازخوردهای کاربران در شبکه‌های اجتماعی و ...، می‌توانند و استراتژی‌های بازاریابی خود را بهبود دهند.

۲. مدیریت ارتباط با مشتریان (CRM): سیستم‌های CRM از تحلیل احساسات برای شناسایی و مدیریت احساسات مشتریان در تعاملات آن‌ها با شرکت استفاده می‌کنند. این فرآیند می‌تواند به شرکت‌ها کمک کند تا مشتریان ناراضی را شناسایی کرده و اقدامات لازم برای بهبود تجربه آن‌ها انجام دهند.

۳. تحلیل سیاسی و اجتماعی: تحلیل احساسات می‌تواند برای شناسایی نگرش عمومی نسبت به مسائل سیاسی یا اجتماعی مورد استفاده قرار گیرد. این روش در پیش‌بینی نتایج انتخابات، تحلیل تأثیرگذاری کمپین‌ها و درک واکنش عمومی به رویدادهای خاص مؤثر است.

۴. صنعت سرگرمی: در صنعت سرگرمی، تحلیل احساسات برای ارزیابی واکنش مخاطبان به فیلم‌ها، سریال‌ها یا موسیقی‌ها به کار می‌رود. این داده‌ها به تولیدکنندگان محتوا کمک می‌کند تا نیازها و ترجیحات مخاطبان را بهتر درک کنند.

۵. حوزه سلامت روان: تحلیل احساسات در تحلیل متون مرتبط با سلامت روان مانند پیام‌ها یا نظرات کاربران در پلتفرم‌های آنلاین به کار می‌رود. این روش می‌تواند به شناسایی علائم افسردگی، اضطراب یا سایر مشکلات روانی کمک کند.

چالش‌ها، محدودیت‌ها و فرصت‌ها

چالش‌ها و محدودیت‌ها:
۱. نویز و غیرساختاریافتگی داده‌ها: داده‌های شبکه‌های اجتماعی معمولاً نویزی، مختصر و غیرساختاریافته هستند. این ویژگی‌ها شناسایی دقیق رویدادها را دشوار می‌کند.

۲. تفاوت زبان و دامنه: رویدادها در دامنه‌های مختلف (مانند سیاسی، اقتصادی یا اجتماعی) و به زبان‌های متفاوت بیان می‌شوند. این تنوع نیازمند مدل‌هایی است که به‌خوبی قابلیت تعمیم داشته باشند.

۳. تشخیص رویدادهای وابسته به زمینه: بسیاری از رویدادها تنها در صورت تحلیل زمینه‌ای که در آن رخ داده‌اند قابل شناسایی هستند. این امر نیازمند استفاده از مدل‌های پیچیده‌تر برای تحلیل زمینه است.

۴. پیچیدگی زبان‌های طبیعی و تنوع ساختارها: زبان‌های طبیعی به دلیل تنوع ساختاری، وجود قواعد نحوی و معنایی پیچیده، و تفاوت‌های فرهنگی و زبانی، یکی از بزرگترین چالش‌ها در استخراج اطلاعات هستند. متون ممکن است شامل جملات بلند، جملات ناقص یا ساختارهای غیرمعمول باشند که تحلیل و استخراج اطلاعات از آن‌ها را دشوار می‌کند. علاوه بر این، وجود زبان‌های کم‌منبع یا زبان‌هایی با ویژگی‌های زبانی خاص (مانند زبان‌های ترکیبی یا با صرف پیچیده) نیازمند توسعه مدل‌ها و ابزارهای اختصاصی است

۵. کمبود داده‌های برچسب‌دار برای آموزش مدل‌ها: مدل‌های یادگیری ماشین و به ویژه یادگیری عمیق برای دستیابی به عملکرد مناسب نیازمند داده‌های برچسب‌دار و با کیفیت هستند. تهیه این داده‌ها معمولاً فرآیندی زمان‌بر و پرهزینه است و در بسیاری از موارد، داده‌های کافی برای زبان‌های کمتر شناخته‌شده یا دامنه‌های خاص در دسترس نیست. این کمبود داده‌ها باعث کاهش دقت مدل‌ها و محدود شدن کاربردهای آن‌ها در محیط‌های واقعی می‌شود.

۶. نویز و عدم دقت در داده‌های واقعی (مانند شبکه‌های اجتماعی): یکی از چالش‌های مهم در استخراج اطلاعات از داده‌های واقعی، وجود نویز بالا و ساختار غیرمنظم در این داده‌ها است. داده‌هایی نظیر پیام‌های شبکه‌های اجتماعی، شامل اشتباهات املایی، اختصارات، عبارات محاوره‌ای و جملات ناقص هستند که تحلیل آن‌ها را دشوار می‌کند. علاوه بر این، داده‌های خبری یا متون غیرساختاریافته نیز ممکن است شامل اطلاعات نادرست یا مبهم باشند که نیازمند تکنیک‌های پیشرفته برای رفع این مشکلات است.

۷. نیاز به منابع محاسباتی بالا: مدل‌های عمیق و پیشرفته‌ای نظیر BERT و GPT برای پردازش و استخراج اطلاعات به منابع محاسباتی بالایی نظیر واحدهای پردازش گرافیکی یا واحدهای پردازش تانسور نیاز دارند. این نیازمندی‌ها می‌تواند هزینه‌های زیادی برای سازمان‌ها ایجاد کرده و استفاده از این مدل‌ها را در محیط‌هایی با منابع محدود دشوار کند. علاوه بر این، زمان پردازش طولانی برای تحلیل داده‌های حجیم، یکی دیگر از محدودیت‌های استفاده از این مدل‌ها محسوب می‌شود.

فرصت‌ها:

۱. مدیریت بحران و امدادرسانی: شناسایی سریع رویدادهای طبیعی یا انسانی از داده‌های خبری و اجتماعی می‌تواند به مدیریت بهتر بحران و بهبود پاسخگویی کمک کند.

۲. پیش‌بینی روندها: تحلیل داده‌های اجتماعی و شناسایی رویدادها می‌تواند به پیش‌بینی روندهای اقتصادی، سیاسی و اجتماعی کمک کند.

۳. ارتقای سیستم‌های پیشنهاددهنده: شناسایی رویدادهای مرتبط با کاربران در شبکه‌های اجتماعی می‌تواند برای بهبود سیستم‌های پیشنهاددهنده مورد استفاده قرار گیرد.

کاربردهای استخراج اطلاعات

استخراج اطلاعات به دلیل قابلیت تبدیل داده‌های غیرساختاریافته به اطلاعات ساختاریافته، کاربردهای گسترده‌ای در حوزه‌های مختلف پیدا کرده است. این فناوری نه تنها باعث بهبود فرایندهای تصمیم‌گیری و مدیریت داده‌ها می‌شود، بلکه در ارتقای دقت و کارایی سیستم‌های تحلیل و پردازش اطلاعات نیز نقش مهمی ایفا می‌کند. در این بخش، به بررسی برخی از کاربردهای اصلی استخراج اطلاعات می‌پردازیم.

- کاربرد در سیستم‌های پرسش و پاسخ و ترجمه ماشینی

استخراج اطلاعات به عنوان یکی از اجزای کلیدی در سیستم‌های پرسش و پاسخ عمل می‌کند. این سیستم‌ها با شناسایی موجودیت‌ها، روابط و مفاهیم کلیدی در متن، امکان پاسخ‌دهی دقیق به سوالات کاربران را فراهم می‌کنند. به عنوان مثال، در سیستم‌های جستجوی اطلاعات، استخراج روابط میان مفاهیم کمک می‌کند تا پاسخ‌هایی با ارتباط مستقیم به پرسش ارائه شود. در حوزه ترجمه ماشینی، استخراج اطلاعات برای شناسایی و تطبیق موجودیت‌ها و روابط معنایی در متن مبدأ و مقصد استفاده می‌شود. این امر به بهبود کیفیت

ترجمه‌های ماشینی، به‌ویژه در متون فنی و تخصصی کمک می‌کند. علاوه بر این، تحلیل ساختار جمله و درک زمینه‌ای متن از طریق استخراج اطلاعات، به سیستم‌های ترجمه ماشینی امکان می‌دهد ترجمه‌های روان‌تر و دقیق‌تری ارائه کنند.

- کاربرد درتحلیل شبکه‌های اجتماعی و بازاریابی

استخراج اطلاعات نقش حیاتی در تحلیل داده‌های شبکه‌های اجتماعی دارد. با استخراج موجودیت‌ها و روابط از پیام‌های کاربران، می‌توان نگرش‌ها و الگوهای رفتاری آن‌ها را شناسایی کرد. این اطلاعات به شرکت‌ها کمک می‌کند تا استراتژی‌های بازاریابی خود را بر اساس نیازها و علایق مخاطبان هدف‌گذاری کنند. در بازاریابی، تحلیل احساسات کاربران از طریق استخراج اطلاعات، ابزار قدرتمندی برای ارزیابی موفقیت کمپین‌های تبلیغاتی و محصولات است. به عنوان مثال، شناسایی نظرات مثبت و منفی در مورد یک محصول خاص می‌تواند به بهبود طراحی و عرضه محصولات جدید کمک کند.

- کاربرد در پزشکی و تحلیل متون علمی

در حوزه پزشکی، استخراج اطلاعات به شناسایی و سازماندهی داده‌های مرتبط با بیماری‌ها، درمان‌ها و داروها کمک می‌کند. برای مثال، از طریق استخراج روابط میان علائم بیماری و درمان‌های موجود، پزشکان می‌توانند تصمیمات دقیق‌تری در ارائه خدمات درمانی اتخاذ کنند. همچنین، در تحقیقات علمی، استخراج اطلاعات به تحلیل داده‌های حجیم در مقالات پژوهشی و کشف دانش جدید کمک می‌کند. این فناوری در سیستم‌های توصیه‌گر پزشکی نیز کاربرد دارد. با تحلیل داده‌های بیماران و استخراج اطلاعات از پرونده‌های پزشکی، می‌توان راهکارهای درمانی مناسب‌تری پیشنهاد داد.

- کاربرد در پیش‌بینی و مدیریت رویدادها در بحران‌ها

یکی از کاربردهای مهم استخراج اطلاعات، شناسایی و پیش‌بینی رویدادهای مرتبط با بحران‌ها است. با تحلیل داده‌های خبری و اجتماعی، می‌توان رویدادهایی نظیر بلایای طبیعی، شیوع بیماری‌ها و بحران‌های انسانی را پیش‌بینی کرده و مدیریت بهتری بر آن‌ها اعمال کرد. برای مثال، در زمان شیوع بیماری‌های همه‌گیر، استخراج اطلاعات از گزارش‌های خبری و پایگاه‌های داده به شناسایی مناطق پرخطر و اطلاع‌رسانی سریع به سازمان‌های امدادی کمک می‌کند. همچنین، در مدیریت بلایای طبیعی مانند زلزله یا سیل، تحلیل داده‌ها می‌تواند به هماهنگی بهتر نیروهای امدادی و تخصیص منابع کمک کند.

کاربردهای استخراج اطلاعات نشان‌دهنده توانمندی بالای این فناوری در بهبود فرآیندهای مختلف در حوزه‌های گوناگون است. از بهبود کیفیت خدمات پزشکی گرفته تا پیش‌بینی بحران‌ها و ارتقای استراتژی‌های بازاریابی، این فناوری به سازمان‌ها و افراد کمک می‌کند تا از داده‌های حجیم و پیچیده به بهترین نحو استفاده کنند. با پیشرفت‌های بیشتر در این حوزه، انتظار می‌رود کاربردهای استخراج اطلاعات به شکل چشمگیری گسترش یابد.

بحث و نتیجه‌گیری

استخراج اطلاعات به عنوان یکی از حوزه‌های اساسی در پردازش زبان طبیعی و یادگیری ماشین، نقشی حیاتی در تحلیل و ساختاردهی داده‌های متنی ایفا می‌کند. این مقاله مروری، تلاش داشت تا با بررسی جامع رویکردها و تکنیک‌های مرتبط با استخراج اطلاعات، چالش‌ها و کاربردهای این حوزه را تبیین کند. در بخش‌های مختلف، مفاهیم پایه، پیشینه تاریخی، روش‌های سنتی و مدرن و کاربردهای متنوع این فناوری بررسی شدند.

از روش‌های سنتی مبتنی بر قواعد و یادگیری ماشین گرفته تا استفاده از مدل‌های عمیق و ترانسفورمرهای پیشرفته، شاهد تحول قابل توجهی در دقت و کارایی سیستم‌های استخراج اطلاعات بوده‌ایم. همچنین، این فناوری کاربردهای گسترده‌ای در حوزه‌هایی نظیر سیستم‌های پرسش و پاسخ، ترجمه ماشینی، تحلیل شبکه‌های اجتماعی، پزشکی، و مدیریت بحران پیدا کرده است. پیشرفت‌های اخیر، امکان پردازش و تحلیل داده‌های حجیم و پیچیده را فراهم کرده و ابزارهایی قدرتمند برای تصمیم‌گیری و کشف دانش ارائه داده است.

با این حال، چالش‌هایی نظیر پیچیدگی زبان‌های طبیعی، کمبود داده‌های برچسب‌دار، و نیاز به منابع محاسباتی بالا همچنان به‌عنوان موانع اصلی پیش روی این حوزه باقی مانده‌اند. رفع این چالش‌ها از طریق توسعه روش‌های نوآورانه نظیر یادگیری انتقالی، استفاده از داده‌های مصنوعی برای آموزش، و بهینه‌سازی مدل‌های محاسباتی، می‌تواند مسیر پیشرفت این حوزه را هموارتر کند.

با گسترش داده‌های متنی در فضای دیجیتال، اهمیت استخراج اطلاعات در آینده بیش از پیش خواهد بود. این فناوری نه تنها به سازمان‌ها و کسب‌وکارها در تحلیل و مدیریت داده‌های خود کمک خواهد کرد، بلکه در حوزه‌هایی نظیر تحلیل اجتماعی، امنیت سایبری، پیش‌بینی روندهای اقتصادی و سیاسی، و توسعه سیستم‌های هوشمند نیز نقشی کلیدی ایفا خواهد کرد.

در نهایت، سرمایه‌گذاری در تحقیق و توسعه این حوزه، از جمله تهیه داده‌های برچسب‌دار، بهبود زیرساخت‌های محاسباتی، و گسترش مدل‌های چندزبانه، می‌تواند به بهبود چشمگیر عملکرد و گسترش کاربردهای استخراج اطلاعات منجر شود. آینده این حوزه پر از فرصت‌هایی برای بهبود کیفیت زندگی، بهینه‌سازی فرآیندها، و افزایش بهره‌وری در مقیاس جهانی خواهد بود.

Abdallah, S., Kannan, S., & Chang, T. (2024). A Survey on Recent Advances in Named Entity Recognition (NER).
Zhang, J., Zhang, X., & Chen, Z. (2019). A Survey of Relation Extraction of Knowledge Graphs.
Imran, M., Castillo, C., Lucas, J., Meier, P., & Vieweg, S. (2013). Event Detection in Social Media: A Survey.
Medhat, W., Hassan, A., & Korashy, H. (2017). Sentiment Analysis Using Lexicon and Machine Learning-Based Approaches: A Survey.
Abdallah, S., Chang, T., & Kannan, S. (2024). A Survey on Recent Advances in Named Entity Recognition (NER).
Etzioni, O., Banko, M., Soderland, S., & Weld, D. S. (2018). A Survey on Open Information Extraction.
Zhang, S., Sun, H., Fang, W., & Han, J. (2022). A Survey on Machine Reading Comprehension Systems.

Techniques, applications and developments of extracting information from text

MohammadReza Bahrani
Master's student, Department of Computer Engineering, Apadana Institute of Higher Education, Shiraz
Haleh Homayouni
Assistant Professor, Faculty Member of the Computer Engineering Group, Apadana Institute of Higher Education, Shiraz
Zahra TasmimGhatei
MSc in Computer Engineering, Apadana Institute of Higher Education, Shiraz

Abstract

Information Extraction is a key area of natural language processing that aims to extract structured data from unstructured texts. This review paper reviews recent advances in various information extraction techniques, including Named Entity Recognition (NER), results, event recognition, and sentiment analysis. It also discusses the challenges associated with large and diverse data sets, the use of training models, and emerging techniques such as Large Language Models (LLMs). Finally, practical applications in this area in knowledge-based systems, social networks, and data analytics tools are highlighted.

Keywords: Information extraction, event recognition, relationship extraction, sentiment analysis, large language models.