
بسم الله الرحمن الرحیم

محققین:

حمیدرضا لطفی نیا

محبوبه صید محمد خانی

استاد گرامی:

جناب آقای دکتر عبدالهی

موضوع پایان نامه:

پردازش زبان طبیعی (NLP)

نام دانشگاه:

علمی کاربردی جهاد دانشگاهی مشهد

دوره تحصیلی:

کاردانی

رشته تحصیلی:

فناوری اطلاعات و ارتباطات **سال تحصیلی:1394**

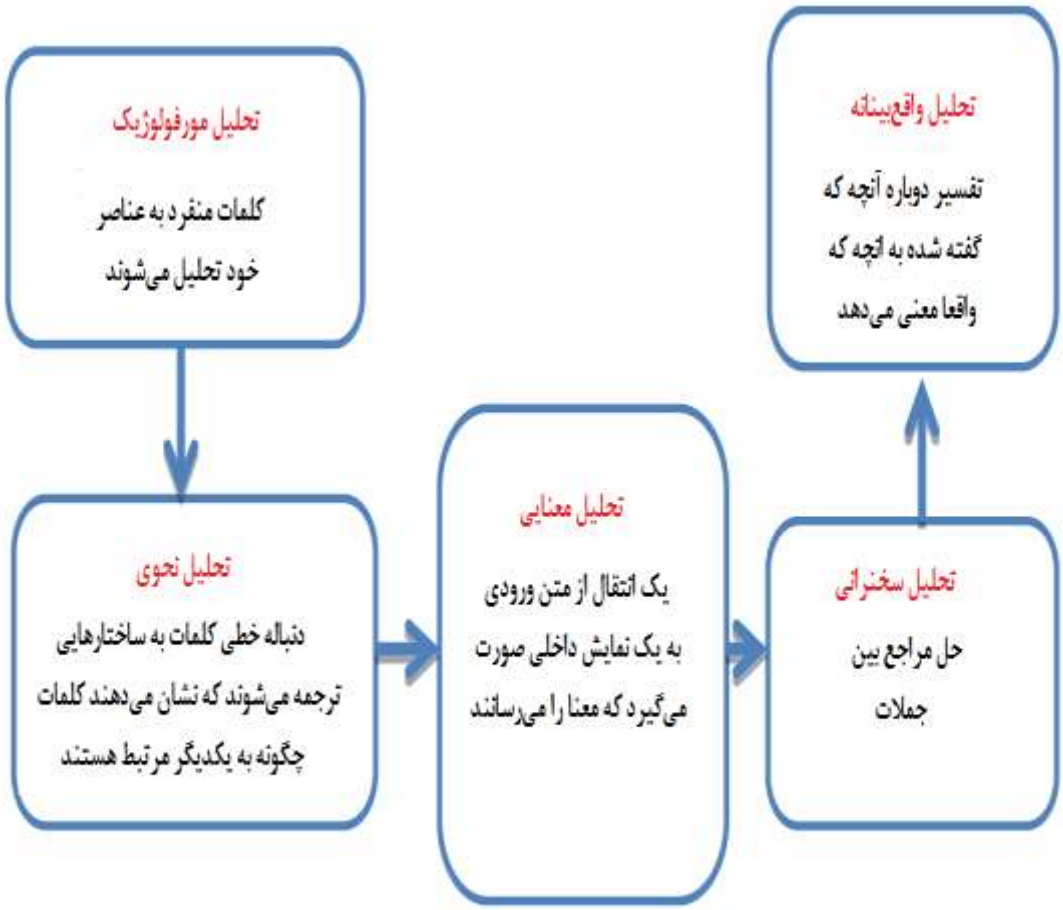
مقدمه

انسان‌ها از توانایی‌های بصری و زبانی خود برای مشاهده اطراف خود و ایجاد ارتباط با یکدیگر استفاده می کنند. آن‌ها با داشتن یک تصویر براحتی می‌توانند یک توصیف زبانی از آن را ارائه داده و بطور مشابه وقتی چیزی بصورت زبانی توصیف شود، می‌توانند تصویری از آن را مجسم کنند. برای انسان‌ها این یک عمل طبیعی است، ولی داشتن این توانایی‌ها در ماشین، یک وظیفه چالش‌برانگیز است. آن به دانش حوزه‌های تحقیقاتی مانند بینایی کامپیوتر و پردازش زبان نیاز دارد. اگرچه، هر دو رشته فوق از هوش مصنوعی گرفته شده و امروزه حوزه‌های تحقیقاتی فعالی را در خود دارند، کار کردن بصورت ایزوله و بدون ارتباط با حوزه دیگر، برای محققان این دو حوزه یک تهدید است. تحقیقات انجام شده در یکی از این دو رشته به نظر می‌رسد که کارایی مناسبی در رشته دیگر نداشته باشد. هر دوی این رشته‌ها بصورت مجزا رونق پیدا کرده و اطلاعات بصری و متنی را جدا از هم استخراج می کنند. اما، هنوز بسیاری از مسائل دنیای واقعی وجود دارند که نیازمند دانش و تخصصی از هر دو رشته هستند، چراکه حضور همزمان داده‌های متنی و بصری، امری طبیعی و چشم‌گیر بوده و براحتی در دسترس می‌باشد؛ برای مثال، زیرنویس در ویدیوها، تصاویر تگ شده در سایت‌های شبکه‌های اجتماعی و غیره. همچنین رشد فاحش داده‌های بصری و متنی در وب و انبارداده‌های خصوصی، نیاز به جست‌و جو، ساماندهی و استخراج این داده‌ها به منظورهای مختلف را ایجاد کرده است. بنابراین، کاوش برای امکان‌سنجی و این که این ادغام دو رشته چگونه می‌تواند انجام شده کاربردهای حاصل شده کجا می‌توانند بکار رفته و مفید باشند، امری مهم و اجباری است.

فهرست مطالب **صفحه**

پردازش زبان طبیعی

حوزه دیگری که در نظر گرفته شده، NLP است. اگر خیلی هسته‌ای نگاه کنیم، NLP به معنای توانمند ساختن کامپیوترها برای استخراج معنا از صحبت‌ها یا متن نوشتاری موجود در یک ورودی زبان طبیعی است. همچنین طیف NLP شامل تولید جملاتی به زبان طبیعی توسط کامپیوترها است که شبیه به جملاتی باشند که ما انسان‌ها بکار می‌بندیم. شکل 2 لسیتی از وظایف را که در یک کاربرد معمول از زبان طبیعی انجام می‌شوند، نشان می‌دهد.



شکل ۲: مراحل پردازش زبان طبیعی (NLP)

محتویات زبان طبیعی

تاریخچه زبان طبیعی

به طور کلی تاریخچه پردازش زبان طبیعی از دهه ۱۹۵۰ میلادی شروع می‌شود. در ۱۹۵۰ آلن تورینگ مقاله معروف خود را که امروزه به عنوان ملاک هوشمندی شناخته می‌شود، منتشر ساخت.

نخستین تلاش‌ها برای ترجمه توسط رایانه ناموفق بودند، به طوری که ناامیدی بنگاه‌های تأمین بودجه پژوهش از این حوزه را نیز در پی داشتند. پس از اولین تلاش‌ها آشکار شد که پیچیدگی زبان بسیار بیشتر از چیزی‌ست که پژوهشگران در ابتدا پنداشته‌بودند. بی‌گمان حوزه‌ای که پس از آن برای استعانت مورد توجه قرار گرفت زبان‌شناسی بود. اما در آن دوران نظریه زبان‌شناسی وجود نداشت که بتواند کمک شایانی به پردازش زبان‌ها بکند. در سال ۱۹۵۷ کتاب ساختارهای نحویاثر نوام چامسکی زبان‌شناس جوان آمریکایی که از آن پس به شناخته‌شده‌ترین چهره زبان‌شناسی نظری تبدیل شد به چاپ رسید از آن پس پردازش زبان با حرکت‌های تازه‌ای دنبال شد اما هرگز قادر به حل کلی مسئله نشد.

محدودیت‌ها

پردازش زبان‌های طبیعی بسیار جذاب است و برای ارتباط بین انسان و ماشین محسوب می‌شود و در صورت عملی شدنش به طور کامل می‌تواند تحولات شگفت‌انگیزی را در پی داشته‌باشد. سیستم‌های قدیمی محدودی مانند SHRDLU که با واژه‌های محدود و مشخصی سر و کار داشتند، بسیار عالی عمل می‌کردند، به طوری که پژوهشگران را به شدت نسبت به این حوزه امیدوار کرده‌بودند. اما در تقابل با چالش‌های جدی‌تر زبانی و پیچیدگی‌ها و ابهام‌های زبان‌ها، این امیدها کم‌رنگ شدند. مسئله پردازش زبان‌های طبیعی معمولاً یک مسئله AI-Complete محسوب می‌شود، چرا که محقق شدن آن به طور کامل مستلزم سطح بالایی از درک جهان خارج و حالات انسان برای ماشین است.

موانع اساسی

رایانه برای آن که بتواند برداشت درستی از جمله‌ای داشته باشد و اطلاعات نهفته در آن جمله را درک کند، گاهی لازم است که برداشتی از معنای کلمات موجود در جمله داشته باشد و تنها آشنایی با دستور زبانکافی نباشد. مثلاً جمله حسن سیب را نخورد برای این‌که کال بود. و جمله حسن سیب را نخورد برای این‌که سیر بود. ساختار دستوریِ کاملاً یکسانی دارند و تشخیص این‌که کلمات «کال» و «سیر» به «حسن» برمی‌گردند یا به «سیب»، بدون داشتن اطلاعات قبلی دربارهٔ ماهیت «حسن» و «سیب» ممکن نیست.

دستورِ هیچ زبانی آن‌قدر دقیق نیست که با استفاده از قواعد دستوری همیشه بتوان به نقش هریک از اجزای جمله‌های آن زبان پی برد.

پردازش زبان‌های طبیعی آمارگرا

پردازش زبان‌های طبیعی به‌شکل آمارگرا عبارت است از استفاده از روش‌های تصادفی، احتمالاتی و آماری برای حل مسائلی مانند آنچه در بالا ذکر شد. به‌ویژه از این روش‌ها برای حل مسائلی استفاده می‌کنند که ناشی از طولانی بودن جملات و زیاد بودن تعداد حالات ممکن برای نقش کلمات هستند. این روش‌ها معمولاً مبتنی بر نمونه‌های متنی و مدل‌های مارکف هستند.

کارکردهای مهم پردازش زبان‌های طبیعی

خلاصه‌سازی خودکار

استخراج اطلاعات

بازیابی اطلاعات

تشخیص گفتار

ویرایش

زبان‌شناسی محاسباتی

خلاصه سازی

خلاصه‌سازی خودکار به عملیاتی گفته می‌شود که به وسیله یک برنامه کامپیوتری، باعث کاهش حجم متن و تولید خلاصه‌ای حاوی مهمترین نکات متن می‌شود.