نخستین کنفرانس بین المللی پردازش خط و زبان فارسی
۱۵ و ۱۶ شهریور ۱۳۹۱
دانشگاه سمنان - دانشکده مهندسی برق و کامپیوتر

معرفی روشی جدید در سیستم‌های پردازش زبان فارسی با استفاده از اصول دستوری

پریسا شیروانی[۱]، محمد اسماعیل زمان[۲] و خشایار یغمایی[۳]
[۱] دانشکده مهندسی برق و کامپیوتر ، دانشگاه سمنان shirvani.parisa@gmail.com
[۲] دانشکده مهندسی برق و کامپیوتر، دانشگاه سمنان mezaman04@yahoo.com
[۳] دانشکده مهندسی برق و کامپیوتر، دانشگاه سمنان khashayar.yaghmaie@gmail.com

چکیده

بازشناسی متون یکی از موضوعات تحقیقاتی در حال رشد در سال‌های اخیر است. تاکنون الگوریتم‌های زیادی به این منظور ارائه و پیشنهاد شده‌اند که بر بازشناسی شبه کلمات یا حروف متمرکز بوده‌اند. در این مقاله از ترکیب دو شاخه علمی پردازش تصاویر و پردازش زبان‌های طبیعی، یک الگوریتم سه مرحله‌ای به منظور بازشناسی متون فارسی بر مبنای بازشناسی جملات فارسی ارائه می‌شود. این روش شامل مراحل استخراج شبه کلمات، ساخت کلمات و سپس جملات بالقوه معنی‌دار و در نهایت استفاده از مدل زبانی بایگرام و چند قاعده گرامری به منظور تشخیص جمله صحیح براساس انطباق با گرامر رایج زبان فارسی می‌باشد. آزمایشات متعدد نشان داد دقت روش ارائه شده برای مرحله استخراج شبه کلمات برابر ۹۲ درصد، برای ساخت کلمات و سپس جملات بالقوه معنی‌دار ۹۸ درصد و برای تشخیص جمله صحیح با استفاده از مدل زبانی بایگرام ۸۰ درصد است.

کلمات کلیدی

بازشناسی متن، فارسی، پردازش زبان‌های طبیعی، تشخیص جملات فارسی، جملات بالقوه معنی‌دار، شبه کلمات، برچسب زنی مولفه‌ها، مدل زبانی بایگرام.

۱- مقدمه

بازشناسی متن[۱] که یکی از محوری‌ترین شاخه‌های بازشناسی الگو[۲] است، با پیشرفت روزافزون رایانه تبدیل به یکی از مهمترین موضوعات مطرح در سال‌های اخیر شده است. در واقع به دلیل ویژگی‌هایی از قبیل انجام سریعتر و دقیق‌تر کارها و خستگی‌ناپذیر بودن رایانه‌ها، انجام امور توسط آن‌ها ترجیح داده می‌شود. بنابراین می‌توان یکی از مهمترین و اصلی‌ترین دلایل گسترش بیشمار تحقیقات در حوزه بازشناسی متون را همین تمایل به انجام کارها توسط رایانه دانست.

تحقیقات در زمینه بازشناسی متون چاپی فارسی و عربی تقریبا از اوایل دهه ۱۹۸۰ آغاز شد [۲و۱]. وجود برخی از ویژگی‌های خاص در نگارش این متون مانند وجود نقاط، علائم، تنوع قلم‌ها و همپوشانی حروف با یکدیگر بازشناسی متون فارسی و عربی را با پیچیدگی‌ها و دشواری‌هایی همراه کرده است.

تاکنون کارهای انجام شده در زمینه بازشناسی متون بر بازشناسی حروف و کلمات متمرکز بوده‌اند [۹-۳]. در دهه‌های اخیر با توجه به کاربردهای گسترده مدل‌های زبان، تحقیقات زیادی برای مدلسازی زبان‌های پرکاربرد جهانی و به خصوص زبان انگلیسی انجام شده است. مدل‌های آماری زبان انگلیسی از سال ۱۹۸۰ در سیستم‌های واقعی به کار رفته‌اند[۱۰]. اما برای مدلسازی زبان فارسی هنوز تحقیقات گسترده و قابل اعتنایی صورت نگرفته است. تقطیع و برچسب دهي نحوي_معنايي داده-هاي نوشتاري يكي از فعاليت‌هاي اصلي در طراحي و ساخت هر دادگان زباني براي استخراج مدل زباني است [۱۱]. براي استخراج مدل زبان فارسي يک بسته نرم افزاري نوشته شده، که در چارچوب فرآيند ماركف صفر تا سه مرحله‌اي، توزيع احتمال مشروط كلمات فارسي را در چهار حالت به دست مي دهد.

در این مقاله از مدل‌سازی آماری زبان[۳] در حوزه پردازش زبان‌های طبیعی[۴] و همچنین چند قاعده گرامری به منظور بازشناسی جملات فارسی استفاده شده است.

با استفاده از یک مدل آماری زبان می‌توان احتمال کلمات بعدی را پیش‌بینی کرد. همچنین مدل‌های زبانی دیگری از جمله مدل‌های توانی[۵][۱۱] و گرامرهای مستقل از متن[۶][۱۲] برای مدل‌سازی زبان‌های طبیعی پیشنهاد شده‌اند، اما در عمل مدل‌های آماری N-gram [۱۳] به علت سادگی پیاده‌سازی ترجیح داده می‌شوند. در الگوریتم پیشنهادی از مدل‌سازی زبان فارسی در جهت پالایش گرامری جملات آن استفاده شده است. برای این منظور ابتدا شبه کلمات از متن مورد پردازش استخراج شده و از ترکیب آن‌ها کلمات و سپس جملات بالقوه معنی‌دار ساخته می‌شوند. در نهایت از مدل زبانی 2-gram یا بایگرام (مدل پنهان مارکوف مرتبه اول) به منظور تشخیص جمله صحیح از میان مجموعه‌ای از جملات بالقوه معنی‌دار استفاده شده است.

در بخش دوم این مقاله به جداسازی و استخراج شبه کلمات از متن مورد پردازش پرداخته می‌شود. در بخش سوم ساخت کلمات بامعنی و جملات بالقوه معنی‌دار از ترکیب شبه کلمات استخراج شده ارائه می‌شود. بخش چهارم به مدل‌سازی زبانی و انتخاب جمله صحیح براساس آن می‌پردازد. در بخش پنجم نتایج حاصل از اعمال مدل زبانی بایگرام ارائه شده و بخش ششم به بحث و تحلیل اختصاص می‌یابد. در نهایت نتیجه‌گیری در بخش هفتم ارائه خواهد شد.

۲- جداسازی شبه کلمات

جداسازی شبه کلمات به عنوان اولین مرحله از مراحل بازشناسی دارای اهمیت ویژه‌ای می‌باشد. در حقیقت به دلیل زنجیروار بودن مراحل بازشناسی، اشتباه در یک مرحله منجر به اشتباهات بیشتر در مراحل بعدی می‌گردد. مسلما این موضوع در مورد استخراج