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

نشانه‌گذاری آماری متون فارسی برای استفاده در موتورهای جستجو

محمدمهدی میردامادی[۱]، علی‌محمد زارع بیدکی[۲] و مهدی رضائیان[۳]
[۱] دانشکده برق و کامپیوتر دانشگاه یزد
mirdamadi@stu.yazduni.ac.ir
[۲] دانشکده برق و کامپیوتر دانشگاه یزد
alizareh@yazduni.ac.ir
[۳] دانشکده برق و کامپیوتر دانشگاه یزد
mrezaeian@yazduni.ac.ir

چکیده

نشانه‌گذاری متن، یکی از فعالیت‌های اصلی در حوزه پردازش زبان‌های طبیعی است. اکثر برنامه‌های پردازش زبان‌های طبیعی به یک پیش‌پردازش برای استخراج کلمات متن و تشخیص نشانه‌ها احتیاج دارند. هدف اصلی و نهایی نشانه‌گذاری، بدست آوردن کلمات معنی‌دار همراه با پیشوندها و پسوندهایشان است. این فعالیت متناسب با زبان‌های طبیعی مختلف، می‌تواند سخت یا آسان باشد. در زبان فارسی با توجه به وجود فاصله و نیم‌فاصله، عدم توجه کاربران به فاصله‌گذاری‌ها و نبود قواعد دقیقی در نوشتن کلمات چند قسمتی، تشخیص و نشانه‌گذاری کلمات چند قسمتی و مرکب، با مشکلات و پیچیدگی‌های خاص خود روبه‌رو است.

در این مقاله برآنیم یک روش آماری برای نشانه‌گذاری متون فارسی جهت استفاده در موتورهای جستجو، ارائه کنیم. برای این منظور از احتمال رخداد دوکلمه‌ای‌های موجود در پیکره استفاده شده است. الگوریتم پیشنهادی شامل ۴ فاز است و با دقت ۸۱/۴٪ به نشانه‌گذاری کلمات متون فارسی می‌پردازد. نتایج آزمایشات نشان دادند این روش می‌تواند با نشانه‌گذاری بهتر کلمات، دقت اطلاعات بازیابی شده در موتور جستجو را بهبود بخشد.

کلمات کلیدی

نشانه‌گذاری، پردازش زبان‌های طبیعی، پیکره، موتور جستجو

۱- مقدمه

با گسترش روزافزون رسانه‌های ذخیره‌سازی الکترونیکی و رسانه‌های ارتباطی، و همچنین پیشرفت سریع علم کامپیوتر و فراگیر شدن آن، امروزه با حجم عظیمی از متون نوشتاری دیجیتال و اسناد الکترونیکی مواجه هستیم [۱]. با گسترش اینگونه اسناد، پردازش اسناد و متون مورد نظر از بین حجم عظیمی از اطلاعات متنی به صورت دستی کاری دشوار و در عمل غیرممکن خواهد بود. از این رو پردازش اتوماتیک متون نوشتاری مورد توجه قرار می‌گیرد، که یکی از موضوعات پردازش زبان‌های طبیعی[۱] است.

برای انجام پردازش اتوماتیک متون نوشتاری به کوچکترین واحد معنی‌دار متن یا کلمات بامفهوم نیاز داریم [4]. کلمات بامفهوم، کلمات ساده، مرکب و یا جمعی هستند که یک مفهوم کلی را می‌رسانند، برای مثال "بین الملل" یک کلمه بامفهوم است. گرچه این کلمه در ظاهر دو کلمه املایی[۲] (به دنباله‌ای از حروف اطلاق می‌شود که دارای معنی هستند) به نظر می‌رسد، اما آن را یک نشانه[۳] در نظر می‌گیریم، زیرا در کل به یک چیز اشاره می‌کند و در کنار هم معنی کامل‌تری می‌دهد. نشانه‌گذاری[۴] به معنی تشخیص و استخراج این نشانه‌ها از متون نوشتاری یا گفتاری می‌باشد، که یکی از مسائل اساسی در پردازش زبان‌های طبیعی است.

نشانه‌گذاری و تشخیص صحیح مرز کلمات و عبارات، در بسیاری از سیستم‌های پردازش زبان طبیعی مانند تشخیص گروه‌های نحوی و پردازش آن‌ها در سیستم‌های ترجمه ماشینی[۵]، استخراج اطلاعات، سیستم پرسش‌وپاسخ[۶]، تشخیص نقش‌های موضوعی، موتورهای جستجو[۷] و غیره نقش کلیدی ایفا می‌کند [۲]. با توجه به این کاربردها نشانه‌گذاری صحیح کلمات می‌تواند موجب بهبود در بازدهی فعالیت‌های ذکر شده باشد.

شاید در ابتدای امر نشانه‌گذاری کلمات امری ساده و آسان به نظر برسد، اما باید به این نکته توجه کرد که حتی در زبان‌هایی مانند فارسی و انگلیسی که از فاصله استفاده می‌کنند هم اگر تنها از فاصله به عنوان جداکننده برای نشانه‌گذاری استفاده شود، نتیجه نهایی خیلی مطلوب نخواهد بود، و باید تکنیکی استفاده شود که بتواند مرز کلمات با مفهوم کامل را به خوبی تشخیص دهد.