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.

با توجه به ساختار زبان فارسی، در این زبان با مشکلات خاص خود مواجه هستیم که در ادامه به برخی از آن‌ها اشاره می‌کنیم.

وجود رسم‌الخط‌های مختلف و سبک‌های نگارش متفاوت در زبان فارسی، باعث شده فاصله، معیار قطعی و دقیقی برای تشخیص مرز کلمه نباشد. به طور کل می‌توان گفت به تعداد افراد جامعه سبک‌های نگارش و رسم‌الخط‌های مختلف وجود دارد. دو نوع فاصله‌ی درون کلمه و برون کلمه در متون نوشتاری زبان فارسی وجود دارد، که در متون تایپی به ترتیب از نیم‌فاصله و فاصله استفاده می‌شود. وجود این دو نوع فاصله‌گذاری باعث شده کلمه‌ای مانند "می رفت" را بتوان به سه صورت "می رفت"، "می‌رفت" و میرفت" نوشت.

اکثر کاربران به فاصله‌گذاری‌ها توجه نمی‌کنند و همچنین قواعد دقیقی در نوشتن کلمات چندقسمتی، وجود ندارد که باعث بروز مشکلات متعددی در نشانه‌گذاری می‌شوند [۳]. همچنین در زبان فارسی میان اشکال ابتدائی، میانی و انتهایی اغلب حروف برحسب موقعیت آن‌ها در یک کلمه تمایز قائل می‌شود. از این روست که وجود یک تا چهار مدل مختلف برای کاراکترها، موجب پیچیدگی در تشخیص برخی کلمات می‌شود. از دیگر مشکلات می‌توان به حروف و کلمات وارد شده از زبان عربی به زبان فارسی اشاره کرد. برای مثال صداهای همزه و تنوین باعث می‌شوند کلمات "پائیز" و "پاییز" ، "حتماً" و "حتما" به دو شکل نوشته شوند. ابهام یونیکد برای حروف "ک" و "ی"، افعال و کلمات مرکب، تولید کلمات جدید و همچنین ورود واژه‌ها از زبان‌های دیگر که در بکارگیری حروف برای نوشتن این کلمات ابهام ایجاد می‌کند، از دیگر مشکلاتی هستند که برای نشانه‌گذاری، با آن‌ها روبه‌رو هستیم.

در این مقاله به دنبال ارائه روشی برای بدست آوردن نشانه‌های صحیح متون فارسی هستیم. ابتدا، در بخش ۲ مروری بر کارهای گذشته‌ای که در این زمینه انجام شده است، داریم. در بخش ۳ به معرفی روش پیشنهادی برای نشانه‌گذاری متون فارسی می‌پردازیم. بخش ۴ شامل نتایج پیاده‌سازی الگوریتم می‌باشد، و در پایان در بخش ۵ نتیجه‌گیری و کارهای آینده آورده شده است.

۲- مروری بر کارهای گذشته

روش‌های گوناگونی برای نشانه‌گذاری کلمات وجود دارند. پرکاربردترین آن‌ها روش‌های مبتنی بر قواعد، روش‌های آماری، روش‌های مبتنی بر فرهنگ واژگان و روش‌های یادگیری هستند. این روش‌ها ممکن است به تنهایی یا به صورت ترکیبی مورد استفاده قرار گیرند.

۲-۱- روش‌های مبتنی بر قواعد

این روش‌ها به دانش زبان شناسی اعم از معنایی و نحوی احتیاج دارند. این قواعد می‌توانند توسط انسان به صورت دستی تعریف شوند و یا از منابع زبانی مانند پیکره‌های برچسب خورده با استفاده از یک رویه یادگیری استخراج شوند.

۲-۲- روش‌های آماری

این روش‌ها به دانش زبان شناسی نیاز ندارند و میزان موفقیت آن‌ها وابستگی زیادی به منابع آمارگیری و پیکره‌ها دارد. روش‌های آماری قابل حمل‌تر از روش‌های دیگر هستند و معمولاً مستقل از زبان‌اند. این روش‌ها، عبارات پر رخداد زبان، فرکانس تکرار و احتمال وقوع عبارات مختلف را به عنوان اطلاعات آماری از منابع زبانی مانند پیکره‌های پردازش شده، اسناد وب، خروجی موتورهای جستجو و غیره استخراج می‌کنند.

۲-۳- روش‌های مبتنی بر فرهنگ واژگان

این روش‌ها با تطبیق کلمات جمله با مدخل‌های یک فرهنگ واژگان، نشانه‌گذاری کلمات را انجام می‌دهند. دقت آن‌ها به پوشش فرهنگ واژگان بستگی دارد و اگر با کلمه‌ی جدیدی روبه‌رو شوند، شکست می‌خورند. در این روش نیاز است، از ابزارهای ریشه‌یابی و تحلیل ساختواژی برای کاهش تعداد کلمات ناموجود در فرهنگ واژگان استفاده شود.

۲-۴- روش‌های یادگیری

در روش‌های یادگیری سیستم، اطلاعات مربوط به نشانه‌گذاری را از منابع ورودی دریافت می‌کند. این اطلاعات می‌توانند مدل زبانی، قواعد نحوی و معنایی و یا اطلاعات آماری مورد نیاز سیستم باشند. در این روش‌ها منبع یادگیری عمدتا پیکره‌ها و واژگان هستند. پیکره‌های برچسب خورده با مقوله نحوی که نشانه‌گذاری شده‌اند یکی از مناسب‌ترین منابع زبانی برای یادگیری قواعد نشانه‌گذاری می‌باشند که در بسیاری زبان‌ها از جمله فارسی در دسترس نمی‌باشند. لذا، در این زبان‌ها، عدم وجود پیکره مناسب، استفاده از این روش را با مشکل روبه‌رو می‌کند.

شمس‌فرد در [5] به معرفی سیستمی به نام STeP-1 پرداخته است. در این سیستم برای نشانه‌گذاری کلمات متون فارسی، از

ترکیب روش‌های مبتنی بر قواعد و مبتنی بر فرهنگ‌واژگان استفاده می‌کند.

در [6] چانگ برای سیستم ترجمه ماشینی از زبان چینی به زبان انگلیسی، به هدف نشانه‌گذاری کلمات چینی، سیستمی را ارائه کرده که از روش‌های آماری و مبتنی بر قواعد توأمان استفاده می‌کند. در سیستم چانگ، اطلاعات در دسترس از پیکره‌های موازی را جهت تشخیص نشانه‌ها برای ترجمه ماشینی ترکیب می‌کند، و از احتمال شرطی و یادگیری بیزی برای بدست آوردن نشانه‌گذار دقیق‌تر استفاده شده است.

فرونزا در [7] یک روش یادگیری باناظر را برای انجام نشانه‌گذاری پیشنهاد می‌کند. در روش او سعی شده عبارات مرکب به عنوان یک نشانه تشخیص داده شود. روش به صورت اتوماتیک، قوانین نشانه‌گذاری را از یک پیکره نشانه‌گذاری شده یاد می‌گیرد. این روش بر روی زبان‌های رومانی و انگلیسی، بهبود قابل توجه آماری دارد.

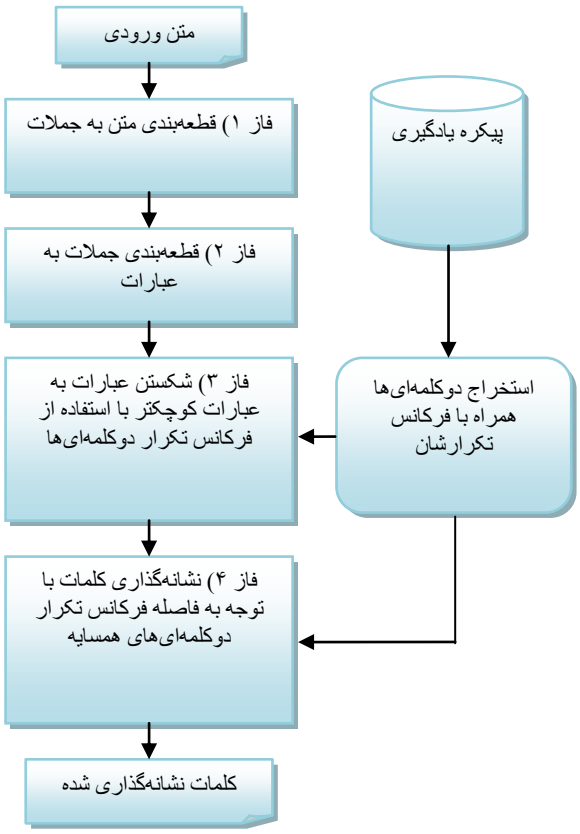
۳- روش ارائه شده

توجه به این نکته ضروری به نظر می‌رسد که نشانه‌گذاری برای موضوعات گوناگون، روند متفاوتی را می‌طلبد. فرض کنید نشانه‌گذاری برای یک سیستم تشخیص گفتار طراحی شده باشد و خروجی آن نشانه‌هایی باشد که برای سیستم مربوطه مناسب باشد، اما ممکن است استفاده از همین نشانه‌گذار برای سیستم ترجمه ماشینی نتیجه مطلوبی به ما ندهد. از این روست که برای هر هدفی باید نشانه‌گذار متناسب با آن هدف طراحی کرد. هدف از ارائه این مقاله، پیشنهاد سیستمی برای نشانه‌گذاری کلمات متون و نوشته‌ها جهت استفاده در موتورهای جستجو می‌باشد. در موتورهای جستجو علاوه بر نشانه‌گذاری کلمات سایت‌ها برای نمایه‌سازی، باید پرس‌وجوی کاربران نیز نشانه‌گذاری شود.

برای نمونه، فرض کنید پرس‌وجوی کاربر "وزیر کشور" باشد. اگر برای نشانه‌گذاری پرس‌وجو تکنیک خاصی بکار گرفته نشود و تنها از فاصله بین کلمات استفاده شود، ممکن است عبارت "وزیر امور خارجه کشور چین" به عنوان یکی از جواب‌های این پرس‌وجو برگردانده شود، که جواب صحیح نمی‌باشد. بنابراین نشانه‌گذاری کلمات برای موتورهای جستجو باید به گونه‌ای انجام شود که کلماتی مثل "وزیر کشور" را به عنوان یک عنصر در نظر بگیرد. برای انجام این کار از احتمال رخداد کلمات در کنار یکدیگر استفاده کرده‌ایم. بدین منظور فرکانس تکرار تمام دوکلمه‌ای‌های همسایه موجود در پیکره را استخراج می‌کنیم، و بر اساس این فرکانس‌ها نشانه‌گذاری را انجام می‌دهیم. در ادامه به توضیح الگوریتم پیشنهادی می‌پردازیم.

۳-۱- الگوریتم پیشنهادی

در ابتدا، پیکره مورد استفاده را به دو قسمت تقسیم می‌کنیم. قسمت اعظم آن را به عنوان داده‌های آموزش در نظر می‌گیریم که در ادامه به نام پیکره یادگیری فراخوانده می‌شود و قسمت کوچکتر به عنوان داده‌های تست و با نام متن ورودی خوانده خواهد شد. هدف نشانه‌گذاری کلمات متن ورودی می‌باشد. برای این منظور الگوریتمی پیشنهاد می‌کنیم که در شکل (۱) شمای کلی آن نشان داده شده است. در ادامه به تشریح هر یک از فازها می‌پردازیم.



شکل (۱): شمای کلی سیستم نشانه‌گذاری کلمات

فاز ۱) در فاز اول جملات متن ورودی را استخراج می‌کنیم. برای این کار از علامت‌هایی که برای پایان جمله استفاده می‌شود، از قبیل "."، "!"، "؟" و غیره استفاده می‌کنیم.

فاز ۲) جمله‌های بدست آمده از فاز قبل را مجدداً به عبارات کوچکتر تقسیم می‌کنیم. برای این کار از حروف اضافه‌ای که بین عبارات در جمله قرار دارند، استفاده می‌شود.

با توجه به اینکه حروف اضافه‌ای چون "و"، "در"، "از" و غیره خود یک نشانه هستند، در جمله، هر یک از این حروف را به

عنوان یک نشانه در نظر گرفته و عباراتی که بین آن‌ها هستند، به فاز بعد منتقل می‌شوند. در رابطه (۱) حرف اضافه ۱و۲ به عنوان نشانه در نظر گرفته می‌شوند و عبارات ۱و۲و۳ به فاز بعد منتقل خواهند شد.

جمله = عبارت ۱ + حرف اضافه ۱ + عبارت ۲
+ حرف اضافه ۲ + عبارت ۳ + نشان پایان جمله

(۱)

فاز ۳) عبارات بدست آمده از فاز ۲ ممکن است حاوی یک کلمه یا بیش‌تر باشند. اگر هر یک از این عبارات شامل یک کلمه بودند، آن کلمه را یک نشانه در نظر می‌گیریم. در غیر اینصورت فرکانس تکرار هر یک از دوکلمه‌ای‌های عبارت، در پیکره یادگیری را با یک حدآستانه مقایسه می‌کنیم. اگر این فرکانس تکرار از حدآستانه بیش‌تر بود، دوکلمه‌ای مورد نظر در یک عبارت قرار می‌گیرد، در غیر اینصورت هر کدام از آن کلمات در عبارتی جدا قرار خواهند گرفت و به اصطلاح عبارت فاز ۲ از آن قسمت شکسته می‌شود.

برای مثال جمله‌ی " آئین نامه خرید کتاب" را در نظر بگیرید. فرض کنید در این جمله دوکلمه‌ای ۱ ="آئین نامه"، دوکلمه‌ای ۲ ="نامه خرید" و دوکلمه‌ای ۳ ="خرید کتاب" باشند.

اگر فرکانس تکرار دوکلمه‌ای ۱ و دوکلمه‌ای ۳ بیش از حدآستانه و فرکانس تکرار دوکلمه‌ای ۲ کمتر باشند، در این صورت عبارت بالا به دو عبارت "آئین نامه" و "خرید کتاب" شکسته شده و به فاز ۴ فرستاده می‌شوند. در این فاز آن دسته از کلماتی که در پیکره یادگیری کمتر در همسایگی هم بوده‌اند و به احتمال کمی تشکیل یک نشانه را می‌دهند، مشخص می‌شوند.

فاز ۴) اگر عبارت حاصل از فاز ۳ شامل یک یا دو کلمه بود، کل عبارت یک نشانه است. برای حالتی که عبارت بیش از دوکلمه داشته باشد، در ابتدا فرض می‌کنیم عبارت از سه کلمه مانند رابطه (۲) تشکیل شده است.

عبارت = کلمه ۱ کلمه ۲ کلمه ۳

(۲)

فرکانس تکرار دوکلمه‌ای‌های "کلمه ۱ کلمه ۲" و "کلمه۲ کلمه ۳" از حدآستانه بیش‌تر است و "کلمه ۲" وجه اشتراک این دوکلمه‌ای‌ها می‌باشد. در صورتی که فرکانس تکرار این دوکلمه‌ای‌ها نزدیک به‌هم باشند، به احتمال زیاد هر سه کلمه یک نشانه‌اند. زیرا این سه کلمه در همسایگی هم، به تعداد زیاد و به نسبت‌های نزدیک به هم در پیکره یادگیری ظاهر شده‌اند.

برای بدست آوردن نسبت تکرار، می‌توان تعداد تکرارشان را بر هم تقسیم کرد. به صورت رابطه (۳):

If [Freq(*word 1* , *word 2*) > Freq(*word 2* , *word 3*)] →
[Freq(*word 1* , *word 2*) / Freq(*word 2* , *word 3*)]
< *Threshold*
Else →
[Freq(*word 2* , *word 3*) / Freq(*word 1* , *word 2*)]
< *Threshold*

(۳)

حاصل این تقسیم را با حدآستانه‌ی دیگری مقایسه کرده و در صورتی که از حدآستانه کوچکتر باشد این سه کلمه، یک نشانه در نظر گرفته می‌شوند. برای عبارات بیش از سه کلمه، ابتدا برای سه کلمه اول همین روال صورت می‌پذیرد، سپس برای کلمات دوم تا چهارم، بعد کلمات سوم تا پنجم و به همین ترتیب به صورت زنجیره‌وار این فرآیند تکرار می‌شود. در هر زنجیره تا جایی که رابطه (۳) بر قرار باشد، کلمات یک نشانه را تشکیل می‌دهند. در انتهای فاز ۴ همه‌ی کلمات متن ورودی نشانه‌گذاری شده‌اند.

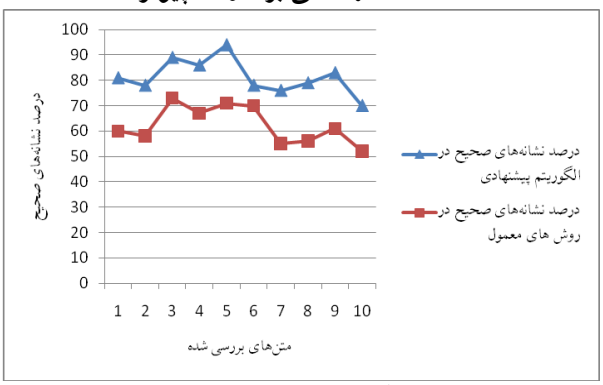
۴- نتایج پیاده‌سازی

برای پیاده‌سازی از پیکره همشهری استفاده گردید. این پیکره مجموعه‌ای از روزنامه‌های همشهری بین سال‌های ۱۹۹۶ تا ۲۰۰۷ میلادی می‌باشد. این مجموعه‌ی عظیم در بیش از ۳۰۰۰ فایل xml ذخیره شده و حاوی ۱۴۵ میلیون کلمه است [10].

جهت پیاده‌سازی الگوریتم پیشنهادی در بخش قبل، پیکره همشهری را به دو قسمت تقسیم کردیم. بیش از ۹۰ درصد از پیکره برای مرحله‌ی آموزش و مابقی برای مرحله‌ی تست در نظر گرفته شد. تمام دوکلمه‌ای‌های قسمت آموزش همراه با فرکانس تکرارشان از پیکره استخراج شدند، که در حدود ۱۸ میلیون دوکلمه‌ای منحصربه‌فرد بودند.

نشانه‌گذاری کلمات قسمت تست با استفاده از روش‌های معمول می‌تواند به طور میانگین ۶۲/۳٪ از کلمات را به درستی نشانه‌گذاری کند که با اعمال الگوریتم پیشنهادی و با استفاده از فرکانس تکرار دوکلمه‌ای‌های بدست آورده، میانگین بازدهی سیستم به ۸۱/۴٪ افزایش یافت. منظور از روش‌های معمول، روش‌هایی است که تنها از فاصله‌ی موجود بین کلمات برای نشانه‌گذاری استفاده کرده و از الگوریتم خاصی پیروی نمی‌کنند.

شکل (۲) مقایسه‌ی بین درصد نشانه‌هایی که در ده متن متفاوت، با استفاده از الگوریتم پیشنهادی و روش‌های معمول، به درستی تشخیص داده شده‌اند را نشان می‌دهد.



شکل (۲): مقایسه الگوریتم پیشنهادی با روش‌های معمول

درصدهای ارائه شده در شکل (۲) با توجه به نشانه‌های مطلوب در موتور جستجو بدست آمده‌اند.

هدف از انجام این پروژه، طراحی یک نشانه‌گذار مؤثر برای استفاده در موتورهای جستجو بود. در موتورهای جستجو مطلوب آن است کلماتی که به یک مفهوم اشاره می‌کنند، یک نشانه در نظر گرفته شوند. برای مثال عبارت "جمهوری اسلامی ایران" یک منظور را می‌رساند، که در سیستم ارائه شده این عبارت به عنوان یک نشانه شناخته می‌شود.

با استخراج فرکانس تکرار دوکلمه‌ای‌های پیکره، به صورت برون‌خطی، و ذخیره داده‌های بدست آمده در جدول هش، سیستم از لحاظ سرعت به خوبی عمل می‌کند.

۵- نتیجه‌گیری و کارهای آینده

در این مقاله اولین روش آماری برای نشانه‌گذاری متون فارسی جهت استفاده در موتورهای جستجو ارائه گردید. این روش با استفاده از احتمال رخداد دوکلمه‌ای‌های موجود در پیکره در چهار فاز انجام گرفت.

در هر فاز تعدادی از نشانه‌ها تشخیص داده شد و نسبت به فازهای قبلی عبارات کوچکتری بدست آمد. در پایان فاز آخر تمام کلمات متن ورودی نشانه‌گذاری شدند. نتایج پیاده‌سازی نشان داد که روش مذکور، کارائی روش‌های معمول را از ۶۲/۳٪ به ۸۱/۴٪ درصد افزایش داد. از دیگر مزایای این روش مستقل از زبان بودن آن است، که می‌تواند برای زبان‌های دیگر از جمله انگلیسی بکار رود.

در سیستم ارائه شده تنها از فرکانس تکرار دوکلمه‌ای‌ها استفاده شد، ممکن است بکارگیری فرکانس تکرار تک کلمه‌ها موجب افزایش کارائی و دقت سیستم شود. به عنوان کارهای آینده می‌توان بر روی این موضوع متمرکز شد. از دیگر پیشنهادات می‌توان به استفاده از وب به عنوان پیکره به جای پیکره همشهری برای مراحل آموزش و تست اشاره کرد. همچنین می‌توان از روش ارائه شده، در موتور جستجوی پارسی‌جو برای مراحل نمایه‌سازی و نشانه‌گذاری کلمات پرس‌وجو، استفاده کرد [11]. این موتور جستجو در دانشگاه یزد طراحی و پیاده‌سازی شده است.

سپاسگزاری

این فعالیت با حمایت‌های موسسه تحقیقات ارتباطات و فناوري اطلاعات انجام گرفته است.

مراجع

[۱] محمدی جنقرا، مسلم، آنالویی، مرتضی، **"استخراج کلمات کلیدی اسناد فارسی"**، سیزدهمین کنفرانس سالانه انجمن کامپیوتر ایران، جزیره کیش، اسفند ۱۳۸۶.

[۲] کیانی، سهیلا، شمس‌فرد، مهرنوش، **"تعیین مرز کلمات و عبارات در متون نوشتاری فارسی"**، چهاردهمین کنفرانس سالانه انجمن کامپیوتر ایران، تهران، اسفند ۱۳۸۷.

[۳] غفوری، سید مجید، راحتی، سعید، پهلوان‌نژاد، محمدرضا، عظیمی‌زاده، علی، **"نرمال‌ساز متون فارسی"**، پانزدهمین کنفرانس بین المللی سالانه انجمن کامپیوتر ایران، تهران، ۱۳۸۸.

[4] Habert, B., Adda, G., Adda-Decker, M., Boula de Marëuil, P., Ferrari, S., Ferret, O., Illouz, G., Paroubek, P., "*Towards Tokenization Evaluation*", Proc. First International Conference on Language Resources and Evaluation (LREC), pp. 427-431, Spain, May 1998.

[5] Shamsfard, M., Kiani, S., Shahedi, Y., "*Step-1: Standard Text Preparation for Persian Language*", proceedings of the 3rd workshop on CAASL-3, MT SUMMIT XII, 2009.

[6] Chung, Tagyoung, Gildea, Daniel, "*Unsupervised tokenization for machine translation*", Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 718–726, Singapore, 2009.

[7] Frunza, Oana, "*A Trainable Tokenizer, solution for multilingual texts and compound expression tokenization*", Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, may 2008.

[8] Graña, Jorge, Alonso, Miguel A., Vilares, Manuel, "*A Common Solution for Tokenization and Part-of-Speech Tagging*", TSD '02 Proceedings of the 5th International Conference on Text, Speech and Dialogue, pp. 3 – 11, London, 2002.

[9] Aleahmad, A., Amiri, H., Rahgozar, M., Oroumchian, F., "*Hamshahri: A standard Persian Text Collection*", Knowledge-Based Systems, vol. 22, no. 5, pp. 382-387, 2009.

[10] Hamshahri Corpus Version 2, http://ece.ut.ac.ir/dbrg/hamshahri/fadownload.html#version2

[11] Parsijoo Search Engine, http://www.parsijoo.ir

زیرنویس‌ها

[1] Natural Language Processing
[2] Orthographic word
[3] Token
[4] Tokenization
[5] Machine Translation Systems
[6] Question Answering System
[7] Search Engines