بهبود برچسب گذاری ادات سخن کلمات ناشناخته‌ی متون فارسی به کمک قوانین انجمنی

محمدحسین الهی‌منش[۱]، بهروز مینایی بیدگلی[۲]

[۱] دانشکده‌ی برق، رایانه و فناوری اطلاعات، دانشگاه آزاد اسلامی واحد قزوین، قزوین
[۱] مرکز تحقیقات کامپیوتری علوم اسلامی نور، قم، ایران
elahimanesh@noornet.net

[۲] استادیار، دانشکده‌ی مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران
[۲] مرکز تحقیقات کامپیوتری علوم اسلامی نور، قم، ایران
b_minaei@iust.ac.ir

چکیده

این مقاله یکی از دغدغه‌های بزرگ زبان‌شناسی محاسباتی[۱] یعنی برچسب‌گذاری ادات سخن[۲] کلمات ناشناخته[۳] را مورد بحث و تحقیق قرار داده است. برچسب‌گذاری ادات سخن که یکی از پایه‌ای‌ترین نیازهای پردازش هوشمند متن به حساب می‌آید، وابسته به زبان متن مورد پردازش است. از این رو فراهم‌سازی برچسب‌گذار با دقت بالا برای زبان فارسی جزو اولویت‌های کار نویسندگان مقاله قرار گرفته است. تکنیک مورد کاربرد ما برای حل مسأله‌ی کلمات ناشناخته، استفاده‌ی ترکیبی از الگوریتم مدل مخفی مارکف[۴] به همراه قوانین انجمنی[۵] بوده است. الگوریتم مدل مخفی مارکف[۶] در بسیاری از برچسب‌گذارهای ادات سخن گذشته به کار گرفته شده[2,3] است و جزو بهترین متدهای مورد استفاده در برچسب‌گذارها به حساب می‌آید. طبق آزمایش‌های انجام شده در این تحقیق، با استفاده از قوانین انجمنی می‌توان دقت برچسب‌گذاری کلمات ناشناخته فارسی را به ۸۱.۲٪ افزایش داد. این در حالی است که میزان دقت کلی و سرجمع برچسب گذار ارائه شده برابر با ۹۸٪ است.

کلمات کلیدی

برچسب‌گذاری ادات سخن، مدل مخفی مارکف، کلمات ناشناخته، قوانین انجمنی.

۱- مقدمه

در سال‌های اخیر مساله پردازش زبان طبیعی[۷] یکی از دغدغه‌های محققین حوزه‌ی کامپیوتر و زبان‌شناسی شده است. استفاده از کامپیوتر و ابزارهای هوشمند باعث شده‌اند که بتوان بسیاری از کارهای مرتبط با متن را با سرعت و دقتی قابل توجه انجام داد. علاوه بر این، قدرت وارد شدن به عرصه‌هایی را که تصور آن‌ها نیز مشکل بوده فراهم کرده است. برای نمونه ترجمه‌ی هوشمند، جستجوگرهای معنایی و بسیاری از کارهای دیگر در این زمینه را می‌توان نام برد. همچنین هر یک از زبان‌های موجود در دنیا به تنهایی می‌تواند مخاطب تمامی پردازش‌های زبانی قرار گیرد. در این راستا و در این مقاله یکی از چالش‌های پیش‌روی پردازش زبان طبیعی، با نام برچسب‌گذاری ادات سخن کلمات ناشناخته، موضوع کار قرار گرفته است. ادامه‌ی این مقاله از ۴ بخش تشکیل شده است. در بخش ۲، مفهوم برچسب‌گذاری ادات سخن ارائه شده است. بخش ۳، کارهای گذشته را ارائه می‌دهد. الگوریتم پیشنهادی در بخش ۴ ارائه شده است. بخش ۵ نتایج حاصل از آزمایش‌ها را بیان می‌کند.

۲- برچسب‌گذاری ادات سخن

در زبان‌شناسی محاسباتی، برچسب‌گذاری ادات سخن (برچسب‌گذاری POS)کلمات یک متن، به فرآیند برچسب‌زنی متناظر هر کلمه با یک برچسب ادات سخن گفته می‌شود. با فرض داشتن رشته‌ی کلمات $\mathbf{W} = \mathbf{w_1}\,\mathbf{w_2}\,\ldots\,\mathbf{w_n}$، مسئله‌ی برچسب‌گذاری ادات سخن این رشته از کلمات را می‌توان به صورت رابطه‌ی زیر نوشت:

$$\mathrm{POS}(\mathrm{W}) = \mathrm{T} \qquad (۱)$$

در رابطه‌ی فوق $\mathbf{T}$ رشته‌ای از برچسب‌های ادات سخن مانند $\mathbf{t_1}\,\mathbf{t_2}\,\ldots\,\mathbf{t_n}$ است.برای نمونه، فرض کنید مجموعه برچسب‌های جدول (۱) در دسترس ماست:

جدول (۱): یک نمونه‌ی ساده از مجموعه برچسب‌های ادات سخن

معادل فارسی	برچسب ادات سخن
فعل	V
اسم	N
ضمیر	PRO
حرف ربط	CONJ
حرف اضافه	P
علائم ویرایشی	PUNC

برچسب‌های معادل کلمات در عبارت «گذشت عمر من اما تو در خیال منی» به صورت زیر خواهد بود:

گذشت	عمر	من	اما	تو	در	خیال	منی
V	N	PRO	CONJ	PRO	P	N	V

۳- کارهای گذشته

در سال‌های اخیر، تحقیقات زیادی در زمینه‌ی برچسب‌گذاری متون فارسی انجام شده است. بسیاری از الگوریتم‌های برچسب‌گذاری ادات سخن را می‌توان در این تحقیقات مشاهده نمود. برای نمونه، الگوریتم مدل مخفی مارکف یکی از پر استفاده‌ترین برچسب‌گذارهای استفاده شده در این تحقیقات است[3,4,5]. نوع دیگری از الگوریتم‌های برچسب‌گذار را برچسب‌گذارهای مبتنی بر حافظه تشکیل می‌دهد. در میان تحقیقات پیشین، رجا و همکارانش از این نوع برچسب‌گذار برای متون فارسی استفاده کرده‌اند [6].

علاوه بر روش‌های فوق، روش‌های ترکیبی برچسب‌گذاری نیز مورد استفاده قرار گرفته است. برای مثال شمس‌فرد و همکاران با استفاده از ترکیب قواعد زبان‌شناسی با برچسب‌گذار دوتایی، سعی در حل برچسب‌گذاری کلمات ناشناخته داشته‌اند[7].

در بسیاری از تحقیقات اشاره شده، مسئله‌ی برچسب‌گذاری کلمات ناشناخته مورد ارزیابی قرار گرفته است. برای نمونه، بهمنش و همکاران در تحقیقات خود نشان دادند، الگوریتم TNT قادر به برچسب‌گذاری ۷۷ درصدی کلمات ناشناخته فارسی است[3]. این الگوریتم، نوعی از برچسب‌گذارهای مدل مخفی مارکف به حساب می‌آید که برای تحلیل کلمات ناشناخته از پسوند آن‌ها استفاده می‌کند[2]. این تکنیک حل مشکل کلمات ناشناخته، پیش از این مورد بررسی قرار گرفته است[1]. همچنین، شمس‌فرد و همکاران نشان دادند، با فرض نرخ تکرار ۱۲ درصدی کلمات ناشناخته، می‌توان این کلمات را با دقتی برابر با ۸۸٪ برچسب‌گذاری نمود [7]. در ادامه و در بخش نتایج، برخی دیگر از تحقیقات گذشته برچسب‌گذاری کلمات ناشناخته‌ی فارسی معرفی شده‌اند.

۴- الگوریتم پیشنهادی

همان‌طور که اشاره شد، برچسب‌گذار مدل مخفی مارکف یکی از بهترین برچسب‌گذارهای ارائه شده‌ی پیش از این است. الگوریتم پیشنهادی این تحقیقات، نوعی خاصی از الگوریتم فوق است که برای حل مسئله‌ی کلمات ناشناخته از قوانین انجمنی استفاده می‌کند. در حقیقت، کلماتی که در فرایند برچسب‌زنی به عنوان ناشناخته تشخیص داده شوند، توسط الگوریتم پیشنهادی مورد تحلیل قرار می‌گیرد. در این تحقیقات، الگوریتم TNT [2] به عنوان الگوریتم پایه‌ی پیشنهادی مورد استفاده قرار گرفته است و قسمت تشخیص برچسب کلمات ناشناخته این الگوریتم با روش ارائه شده در بخش بعد جایگزین شده است.

۴-۱- حل مسئله کلمات ناشناخته

کلمه‌ی ناشناخته، کلمه‌ای است که رخداد آن از یک آستانه‌ی خاص در پیکره‌ی آموزشی روش‌های آماری پایین‌تر بوده و یا موتورهای قاعده محور قابلیت پردازش آن را ندارند. در تحقیقات گذشته، این آستانه مقادیری بین ۰ تا ۱۰ را داشته است. کلمه‌ی ناشناخته با نام کلمه‌ی نادیده[8] و کلمه‌ی خارج از لغت‌نامه[9] نیز یاد می‌شود.

با توجه به قدرت زایایی زبان طبیعی برای تولید کلمات جدید، همواره مشکل کلمات ناشناخته گریبان‌گیر مسائل پردازش زبان طبیعی بوده است. از این رو سلسله مسائل شکل گرفته برای این نوع از کلمات از شناسایی آن‌ها[8][10]شروع شده و تا مراحل تحلیل نحوی[9][11]، معنایی[10][12] و نقش واژه‌ای[11][13] نیز ادامه پیدا کرده است.

در این تحقیق، کلمه‌ی ناشناخته در دو مرحله تحلیل شده است. در مرحله‌ی اول، چند روش ابداعی[14] ساده برای تشخیص برچسب کلمه‌ی ناشناخته به کار رفته است. مرحله‌ی دوم، کلماتی را که توسط مرحله‌ی اول تحلیل نشده‌اند مورد پردازش خود قرار می‌دهد. در نهایت، و در صورت عدم توانایی دو مرحله‌ی فوق در تحلیل کلمه‌ای خاص، این کلمه با برچسب N که پراحتمال‌ترین برچسب برای کلمات ناشناخته است برچسب می‌خورد. این برچسب به معنای اسم است. در ادامه، دو مرحله‌ی یاد شده‌ی فوق به طور مفصّل ارائه شده است.

۴-۱-۱- چند روش ابداعی ساده

بسیاری از کلمات ناشناخته، از الگوهای ساده‌ای پیروی می‌کنند. برای نمونه می‌توان اعداد را نام برد. کلمه‌های به فرم عدد، شامل کاراکترهای ۰ تا ۹ به همراه کاراکتر جدا کننده اعشار (.) می‌شوند. در پیکره‌ی مورد آزمایش این تحقیق، اعداد برچسب خاص خود با نام NUM را دارند. تشخیص این نوع کلمات توسط یک عبارت منظم[15] امکان‌پذیر است.

علاوه بر اعداد، نوع دیگر از کلمات، شامل بخشی عددی و بخشی از سایر کارکترهای فارسی می‌شوند. این نوع از کلمات، غالبا برچسب ADJ به معنای صفت را دارند. برای مثال کلمه «۲۲قسمتی» را می‌توان نام برد. تشخیص این نوع از کلمات نیز توسط عبارات منظم امکان‌پذیر است.

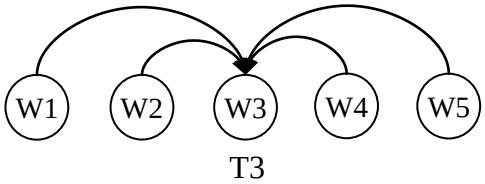
نوع سوم ابداع استفاده شده، برای کلمات انگلیسی است. این گونه کلمات برچسب RES می‌خورند که به معنای متفرقه است. این کلمات

از حروف الفبای انگلیسی به همراه اعداد تشکیل می‌شوند که حضور حداقل یکی از حروف الفبای انگلیسی در آن‌ها حتمی است. برای تشخیص این نوع از کلمات نیز از عبارات منظم استفاده شده است.

۴-۱-۲- ساخت قوانین انجمنی

مرحله‌ی دوم تحلیل کلمات ناشناخته، استفاده از قوانین انجمنی است که در این بخش به آن می‌پردازیم. با این فرض که هر کلمه‌ی موجود در پیکره‌ی آموزشی، یک تراکنش مسأله‌ی قوانین انجمنی است. و با دانستن برچسب متناظر هر یک از این کلمات، می‌توان قوانینی را تولید نمود که قسمت شرط آن‌ها ویژگی‌ای از کلمه‌ی مورد نظر و قسمت نتیجه، برچسب آن کلمه باشد.

هر کلمه و همسایه‌های آن را می‌توان به صورت گراف جهت‌دار زیر در نظر گرفت.



شکل (۱): گراف جهت‌دار تاثیر کلمات پیرامون یک کلمه بر آن کلمه

گراف فوق، کلمه‌ی W3 و دو همسایه‌ی چپ و دو همسایه‌ی راست آن را نشان می‌دهد. در این گراف، برچسب کلمه‌ی W3 با T3 مشخص شده است. می‌توان نشان داد، شعاع همسایگی ۲ برای تاثیر کلمات مجاور یک کلمه بر آن کلمه از لحاظ هزینه‌ی پردازشی و بهبود تحلیل‌های کلمه‌ی مورد نظر، بهترین حالت است. جدول زیر، شامل مجموعه قوانینی است که قسمت نتیجه‌ی آن‌ها برچسب کلمه‌ی W3 (T3) بوده و قسمت شرط آن‌ها از گراف فوق حاصل می‌شود.

جدول (۲): قوانین انجمنی حاصل از همسایه‌های کلمه‌ای مانند W3.

شماره	قسمت شرط	قسمت نتیجه
۱	W1W2W4W5	T3
۲	W1W2W4-	T3
۳	W1W2-W5	T3
۴	W1W2--	T3
۵	W1-W4W5	T3
۶	W1-W4-	T3
۷	W1--W5	T3
۸	W1---	T3
۹	-W2W4W5	T3
۱۰	-W2W4-	T3
۱۱	-W2-W5	T3
۱۲	-W2--	T3
۱۳	--W4W5	T3
۱۴	--W4-	T3
۱۵	---W5	T3
۱۶	----	T3

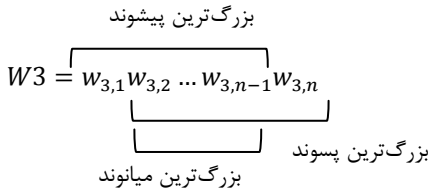
برای تولید جدول فوق، از ۴ کلمه‌ی پیرامون کلمه W3 استفاده شده است. هر کلمه در بخش شرط، یکی از دو حالت حضور یا عدم حضور را می‌تواند داشته باشد. عدم حضور کلمه‌ای مانند W1 با علامت - مشخص شده است.

نوع دیگر از قوانین انجمنی را می‌توان از خود کلمه‌ی W3 (صرف نظر از ویژگی‌های زمینه‌ی این کلمه) به دست آورد. اگر فرض کنیم طول کلمه‌ی W3 برابر با n باشد، می‌توان این کلمه را به صورت زیر نشان داد:

$$\boldsymbol{W3 = w_{3,1}w_{3,2} \dots w_{3,n-1}w_{3,n}} \quad (۲)$$

در رابطه‌ی فوق، هر یک از $w_{3,i}$ ما یک کاراکتر از کلمه‌ی W3 را نشان می‌دهد. با این فرض که بزرگ‌ترین پسوند کلمه‌ی W3 از $w_{3,2}$ شروع شده و تا$w_{3,n}$ ادامه پیدا می‌کند، تعداد پسوندهای ممکن این کلمه برابر با n-1 است. در این فرض، پسوندهای زبان‌شناسی کلمه‌ی W3 در نظر گرفته نشده و در حقیقت ترتیبی از چند کاراکتر انتهایی این کلمه به عنوان پسوند در نظر گرفته شده است.

با همین نوع از تولید، بزرگ‌ترین پسوند را می‌توان برای بزرگ‌ترین پیشوند و بزرگ‌ترین میانوند کلمه‌ی W3 داشت. شکل زیر، بزرگ‌ترین وندهای ممکن کلمه‌ی W3 را که توسط روش ما تولید شده، نشان داده است.



شکل (۲): بزرگ‌ترین پیشوند، میانوند و پسوند کلمه W3.

با داشتن بزرگ‌ترین وندها و نحوه‌ی محاسبه این وندها می‌توان تمامی پیشوندها، میانوندها و پسوندهای کلمه‌ی W3 را تولید نمود. جدول زیر تعداد پیشوندها، میانوندها و پسوندهای ممکنی را که توسط این روش برای W3 تولید می‌شود نشان می‌دهد.

جدول (۳): تعداد وندهای تولید شده روش ما برای کلمه‌ی W3.

نوع وند	تعداد
پیشوند	n-1
میانوند	$\frac{(n-2)(n-1)}{2}$
پسوند	n-1

ما برای تولید قوانین انجمنی، به ازای هر کلمه‌ی پیکره آموزشی، مجموعه قوانین حاصل از زمینه و خود کلمه را محاسبه کرده‌ایم. تعداد قوانین تولید شده‌ی هر کلمه به صورت زیر محاسبه می‌شود:

$$|\boldsymbol{R_{w3}}| = \frac{\boldsymbol{1}}{\boldsymbol{2}} * (\boldsymbol{n^2 + n}) + \boldsymbol{16} \quad (۳)$$

برای مثال، این مقدار برای کلمه‌ای با طول ۵، برابر ۳۱ قاعده است. تعداد قوانین تولید شده و توزیع‌های مختلف این قوانین در بخش نتایج آزمایش‌ها ارائه شده است. نتیجه فرایند ساخت قوانین انجمنی، تولید چهار مجموعه قوانین[۱۶] است که به ترتیب مجموعه‌ی قوانین زمینه، مجموعه‌ی قوانین پسوندی، مجموعه‌ی قوانین میانوندی و مجموعه‌ی قوانین پیشوندی نام دارند.

۴-۱-۳- ارزیابی قوانین انجمنی

دو معیار مرسوم ارزیابی قوانین انجمنی، معیار پشتیبانی[۱۷] و معیار اطمینان[۱۸] است. با فرض داشتن قاعده‌ی R مانند زیر:

$$\boldsymbol{R: A \rightarrow B} \qquad (۴)$$

میزان معیار پشتیبانی قاعده R برابر با نسبت تعداد تراکنش‌های شامل بخش A به تعداد کل تراکنش‌ها است. اگر فرض کنیم S مجموعه‌ی کل تراکنش‌ها باشد، میزان این معیار که با sup(R) نشان داده شده، به صورت زیر محاسبه می‌گردد:

$$\boldsymbol{\sup(R) = \frac{|(s \in S;\ s\ contain\ A \,\&\, B)|}{|S|}} \qquad (۵)$$

در رابطه فوق، |S| برابر با تعداد عناصر مجموعه‌ی S است. این نکته را یادآور می‌شویم که صورت رابطه‌ی فوق، در ادامه با نام تعداد پشتیبانی مورد استفاده قرار می‌گیرد.

معیار اطمینان قاعده R برابر با نسبت میزان رخداد B در تراکنش‌های شامل A به تعداد تراکنش‌های شامل A است. این میزان که با conf(R) نشان داده شده، به صورت زیر محاسبه می‌گردد:

$$\boldsymbol{\mathrm{conf}(R) = \frac{|(s \in S;\ s\ contain\ A \,\&\, B)|}{|(s \in S;\ s\ contain\ A)|}} \qquad (۶)$$

معیارهای دیگری نیز همچون جذابیت[۱۹]، جاکارد[۲۰] و ضریب فی[۲۱] برای ارزیابی قوانین انجمنی استفاده شده که در این تحقیق به کار نرفته‌اند. برای اطلاعات بیشتر می‌توانید به [12] مراجعه کنید.

ارزیابی قوانین انجمنی تولید شده از بخش قبل با توجه به پیکره‌ی آموزشی انجام می‌گیرد. هر یک از کلمات پیکره‌ی آموزشی به عنوان یک تراکنش در نظر گرفته شده است.

۴-۱-۴- استفاده از قوانین انجمنی

قوانین حاصل از تولید قوانین انجمنی، پس از ارزیابی، فیلتر شده و مجموعه قوانین ما را می‌سازد. این مجموعه قوانین، پایه بخش تحلیل کلمات ناشناخته‌ی الگوریتم پیشنهادی را تشکیل می‌دهد. کلماتی که طی فرایند برچسب‌گذاری، ناشناخته تشخیص داده شده و توسط مرحله‌ی اول تحلیل کلمات ناشناخته تحلیل نشده‌اند، به کمک این قوانین تحلیل می‌شوند.

برای این کار، ابتدا فهرست تمامی بخش‌های شرط قوانین انجمنی که از کلمه‌ی ناشناخته مورد نظر حاصل پذیر است، تولید شده. حال، تمامی قوانینی از مجموعه‌ی قوانین که شرط‌های فهرست فوق را شامل می‌شود آماده می‌گردد. فهرست حاصل از این کار، به عنوان فهرست قوانین فعال شده برای کلمه‌ی ناشناخته‌ی مورد نظر به حساب می‌آید. این فهرست در تحقیقات ما، با نشانه‌ی L مشخص شده است.

با توجه به اینکه فهرست L جواب‌های مختلفی برای برچسب‌زنی کلمه‌ی ناشناخته در خود دارد، نیاز است تا نتایج آن را بر حسب اولویت مرتب کنیم. برای این کار، فهرست L را به ترتیب بر حسب اطمینان و سپس پشتیبانی مرتب می‌کنیم. با توجه به این که این اولویت بندی، به هر دو معیار اطمینان و پشتیبانی اهمیت داده، نتایج خوبی را نسبت به حالت استفاده‌ی تکی از اطمینان و یا پشتیبانی در برخواهد داشت. فهرست مرتب شده‌ی فوق را SL می‌نامیم که جواب اول آن می‌تواند به عنوان برچسب کلمه‌ی ناشناخته استفاده گردد. برای مثال، اگر کلمه‌ی مورد پرسش «انسان گرا» باشد، چند عنصر ابتدای فهرست SL برابر است با:

جدول (۴): فهرست هفت قانون برتر فعال شده برای کلمه «انسان گرا»

شماره	نوع قانون	شرط	نتیجه	تعداد پشتیبانی	اطمینان
۱	پسوندی	ن گرا	ADJ	۲۲	۱
۲	پسوندی	ان گرا	ADJ	۹	۱
۳	پیشوندی	نسان گر	N	۷	۱
۴	پسوندی	سان گر	N	۷	۱
۵	پیشوندی	انسان گر	N	۷	۱
۶	پسوندی	ـگرا	ADJ	۷	۰.۹۶
۷	پسوندی	گرا	ADJ	۲۱۷	۰.۹۴

همان‌طور که جدول (۴) نشان می‌دهد، اولین قانون، کلمه‌ی «انسان گرا» را با ADJ برچسب می‌زند. برچسب صحیح این کلمه ADJ است که این مطلب نشان از تشخیص صحیح قانون ۱ دارد.

مثالی دیگر از کاربرد قوانین انجمنی، کلمه‌ی «محسوسش» است چند عنصر ابتدایی فهرست SL متناظر این کلمه در جدول زیر قرار داده شده است:

جدول (۵): هفت قانون برتر فعال شده برای کلمه‌ی «محسوسش»

شماره	نوع قانون	شرط	نتیجه	تعداد پشتوانه	اطمینان
۱	پسوندی	وسش	N	۵	۱
۲	میانوندی	حسو	ADJ	۱۳۵۰	۰.۹۸
۳	پیشوندی	محسو	ADJ	۱۳۴۲	۰.۹۸
۴	پسوندی	سش	N	۵۶۹	۰.۹۷
۵	میانوندی	سوس	N	۵۸۷	۰.۸۴
۶	پسوندی	ش	N	۹۷۲۰۶	۰.۸
۷	میانوندی	حسوس	ADJ	۵۰	۰.۷۹

همان‌طور که جدول (۵) نشان می‌دهد، اولین قانون، کلمه‌ی «محسوسش» را با N برچسب می‌زند. برچسب صحیح این کلمه ADJ

است که این نشان از اشتباه کردن قاعده ۱ دارد. این اشتباه می‌تواند به دو طریق اصلاح گردد. راه اول، استفاده از تکنیک بهتری برای مرتب کردن قوانین است. راه دوم که مورد استفاده این الگوریتم بوده، استفاده‌ی فازی از تمامی جواب‌ها است. جواب‌های دارای رتبه‌ی بالاتر، نسبت تعلق بیشتری را برای برچسب زنی کلمه‌ی ناشناخته با قسمت نتیجه‌ی آن‌ها دارند. جواب‌های فازی[۲۲] حاصل از این کار توسط برچسب‌گذار رفع ابهام می‌گردد.

۴-۱-۵- استفاده از جواب‌های بعدی قوانین انجمنی

برچسب‌گذار مدل مخفی مارکف نیاز به جدول احتمال برچسب‌ها به ازای هر کلمه دارد. هدف از فازی سازی جواب‌های ارائه شده فهرست SL، ساخت جدول احتمال مربوط به کلمه‌های ناشناخته است. رابطه‌ی زیر، نحوه‌ی محاسبه‌ی مقدار فازی تعلق برچسب تخمین زده‌شده T به ازای کلمه‌ی ناشناخته $W_{unknown}$ را نشان می‌دهد:

$$P(W_{unknown}|T) = \frac{1}{1+\alpha * \beta}(1 - P(T)) \qquad (۷)$$

در رابطه‌ی فوق، $P(T)$ احتمال پیشین[۲۳] برچسب T است. این احتمال برابر با نسبت تعداد رخداد آن برچسب در پیکره‌ی آموزشی به کل کلمات پیکره‌ی آموزشی است. مقدار α مشخص می‌کند که برچسب T چندمین پیشنهاد مجموعه قوانین است. مقدار β نیز ضریبی ساده، مابین ۱-۲۰ است که طی آزمایش‌ها تنظیم می‌گردد.

۵- نتایج آزمایش‌ها

آزمایش‌های این تحقیق، روی پیکره‌ی ۱۰ میلیونی بی‌جن‌خان [13] انجام شده است. کلمات این پیکره، در دو سطح (در یک سطح با ۱۴ برچسب اصلی و در سطح دیگر با ۶۰۶ برچسب جزئی) برچسب خورده است. این نسخه‌ی پیکره‌ی بی‌جن‌خان، جدیدترین نسخه‌ی این پیکره به حساب می‌آید. طی آزمایش‌های این مقاله، ۹۰٪ داده‌ی پیکره‌ی فوق به عنوان داده‌ی آموزشی و مابقی آن به عنوان داده‌ی آزمایشی در نظر گرفته شده است. در این تحقیقات، سطح اول برچسب‌های این پیکره مورد استفاده قرار گرفته است. نرخ تکرار کلمه‌های شناخته شده و ناشناخته‌ی دادگان آزمایش این مقاله در جدول (۷) نمایش داده شده است. مقادیر محاسبه شده در این جدول با فرض آستانه ناشناختگی ۰ در نظر گرفته شده است.

جدول (۶): نرخ تکرار کلمات شناخته شده و ناشناخته در دادگان آزمایش.

درصد پوشش	تعداد رخداد	نوع کلمه
٪۹۹.۲	۹۷۷۴۴۷	شناخته شده
٪۰.۸	۷۵۴۶	ناشناخته

برای ساخت چهار مجموعه قانون تحلیل کلمات ناشناخته، از داده‌ی آموزش استفاده شده است. پس از تولید تمامی قوانین ممکن از روی دادگان آموزشی، قوانینی از پیشوند، میانوند و پسوندها که اطمینانی کمتر از ۰.۶ داشته‌اند حذف شده‌اند. قوانین تشکیل شده از زمینه‌ی کلمه بسیار زیاد هستند. تعداد این قوانین در تحقیق ما بیش از ۵۰ میلیون قانون بوده است. این قوانین ابتدا بر حسب تعداد پشتوانه‌ی کمتر از ۵ فیلتر شده و سپس بر حسب مقادیر اطمینان کمتر از ۰.۶ دوباره فیلتر شده است. جدول زیر، هر مجموعه قانون و تعداد قانون باقی مانده را نشان می‌دهد:

جدول (۷): مجموعه‌های قوانین حاصل از تولید قوانین انجمنی

تعداد قانون	مجموعه قانون
۲۲۸۷۵۷	قوانین پیشوندی
۳۹۱۵۳۲	قوانین میانوندی
۲۵۰۱۵۵	قوانین پسوندی
۱۱۳۶۹۰۵	قوانین زمینه
۲۰۰۷۳۴۹	کل قوانین

جدول زیر نتایج استفاده از مراحل حل مشکل کلمات ناشناخته در این تحقیق را نشان می‌دهد. در حالت پایه، هر کلمه‌ی ناشناخته با برچسب N که محتمل‌ترین برچسب این گونه کلمات است، برچسب خورده است.

جدول (۸) : نتایج ارزیابی برچسب‌گذار پیشنهادی

شماره	روش	دقت برای کلمات شناخته شده	دقت برای کلمات ناشناخته	دقت کلی
۱	حالت پایه	٪۹۸.۱	٪۵۹.۴	٪۹۷.۸
۲	استفاده از ابداع‌ها	٪۹۸.۱	٪۶۶.۲	٪۹۷.۹
۳	استفاده از ابداع‌ها به همراه قوانین انجمنی	٪۹۸.۱	٪۷۹.۶	٪۹۸
۴	استفاده از ابداع‌ها به همراه حالت فازی خروجی قوانین انجمنی	٪۹۸.۱	٪۸۱.۲	٪۹۸

همان‌طور که جدول فوق نشان می‌دهد، روش شماره‌ی ۴، بهترین نتایج را برای برچسب‌گذاری کلمات ناشناخته در بر داشته است. این روش با اختلاف ۲۱.۸٪ نسبت به حالت پایه، دقت ۸۱.۲٪ را برای برچسب‌گذاری کلمات ناشناخته نشان می‌دهد. در این روش مقدار ۱۴ برای β در رابطه (۷) در نظر گرفته شده است.

برای مقایسه‌پذیری روش پیشنهادی با تحقیقات گذشته، با توجه به نرخ‌های متفاوت کلمات ناشناخته در این تحقیقات، سعی شده تا برچسب‌گذار ارائه شده را با همان میزان از کلمات ناشناخته مورد ارزیابی قرار دهیم. برای این کار، کلمات پیکره‌ی آموزشی را بر حسب تعداد رخدادشان در پیکره‌ی آموزشی و به صورت یک فهرست صعودی، مرتب می‌کنیم. سپس آن قدر عناصر ابتدای این فهرست را به عنوان کلمه‌ی ناشناخته انتخاب می‌کنیم تا میزان درصد ناشناختگی مورد نظر حاصل گردد. نتایج مقایسه‌های انجام شده در جدول زیر قابل مشاهده است:

جدول (۹) : نتایج مقایسه الگوریتم پیشنهادی با برخی از بهترین کارهای گذشته

روش	درصد کلمات ناشناخته	دقت کلمات ناشناخته	دقت کلی	نتایج روش پیشنهادی برای کلمات ناشناخته
A1	٪۲	٪۷۳.۵	٪۹۵.۹	٪۸۷.۶
A2	٪۱۲	٪۸۸	٪۹۰.۹	٪۸۹.۵
A3	٪۱.۸	٪۷۹.۴۴	٪۹۶.۹۴	٪۸۷.۳
A4	٪۲	٪۶۹.۳۵	٪۹۶.۰۷	٪۸۷.۶

در جدول فوق، روش A1 توسط محسنی و همکاران و در سال ۱۳۸۷ ارائه شده است [14]. پیکره‌ی مورد استفاده‌ی آزمایش‌های این تحقیق، همان پیکره‌ی مورد استفاده ما بوده است. اما با توجه به جداسازی ۸۰٪ به ۲۰٪ برای دادگان آموزشی و دادگان آزمایشی، نرخ رخداد کلمه‌های ناشناخته این تحقیق، متفاوت از نتایج ما به دست آمده است. روش A2، روش پیشنهادی شمس‌فرد و همکاران را نشان می‌دهد [7]. روش A3 یکی دیگر از تحقیق‌های گذشته است که توسط رجا و همکاران ارائه شده است. میزان نرخ کلمات ناشناخته این تحقیق برابر با ۱.۸٪ اعلام شده است [6]. آخرین روش مورد مقایسه، روش A4 بوده است. این روش که توسط بهمنش و همکاران مورد استفاده قرار گرفته است، میزان ۲٪ را برای نرخ رخداد کلمات ناشناخته اعلام کرده است[3].

در انتها، این نکته را یادآور می‌شویم که قوانین انجمنی، قابلیت بسیار بالایی برای تشخیص برچسب کلمات ناشناخته دارند. در تحقیقات ما و برای پیچیده‌ترین کلمات، به طور میانگین ۱۸ قانون فعال شده است. با توجه به مرتب‌سازی ارائه شده‌ی بخش روش پیشنهادی، احتمال وجود جواب صحیح در جواب اول ۷۹٪، در یکی از دو جواب اول ۹۲٪ و در یکی از سه جواب اول ۹۵٪ است.

سپاسگزاری

در پایان لازم می‌دانیم تا از پژوهشکده متن‌کاوی نور وابسته به مرکز تحقیقات کامپیوتری علوم اسلامی (www.noornet.net) که حامی این تحقیقات بوده‌اند تشکر نماییم.

مراجع

[1] Samuelsson C., *Morphological tagging based entirely on Bayesian inference*, In 9th Nordic Conference on Computational Linguistic NODALIDA-93, Stockholm University, Stockholm, Sweden, 1993.

[2] Brants, T..*TnT: A statistical part of speech tagger*, Proceedings of the 6th Conference on Applied Natural Language Processing, Apr. 29-May 04, Association for Computational Linguistics Morristown, USA. 2000.

[3] Behmanesh, A. A., *Statistical part of speech tagger for Persian words*, ----, 2011

[4] Seraji M., *A statistical part-of-speech tagger for Persian*. InProceedings of the 18th Nordic Conference of Computational Linguistics NODALIDA 2011. NEALT Proceedings Series, pages 340_343, 2011.

[5] Okhovvat, M., Minaei Bidgoli, B, *A hidden Markov model for Persian part-of-speech tagging*. In Proceedings of Procedia CS, 977-981, 2011.

[6] Raja, F., Tasharofi, S. and Oroumchian, F., *Statistical POS tagging experiments on Persian text*, Second Workshop on Computational Approaches to Arabic Script-based Languages, 21-22 July, 2007. Stanford, California, (2007).

[7] Fadaei H., Shamsfard M., *Persian POS Tagging Using Probabilistic Morphological Analysis*, International Journal of Computer Application in Technology (IJCAT), pp. 264-273, 2010.

[8] Guohong Fu, and Kang-Kwong Luke, *Chinese unknown word identification using class-based LM*, Lecture Notes in Artificial Intelligence (IJCNLP 2004), 2005.

[9] Ebach G., *Syntactic processing of unknown words*, IWBS Report 131, IBM, Stuttgart, 1990.

[10] Taylor, J. M., V. Raskin, and C. F. Hempelmann, *Towards computational guessing of unknown word meanings: The ontological semantic approach*, Cognitive Science Conference, Boston, MA, 2011.

[11] Erk K., *Unknown word sense detection as outlierdetection*, In *Proceedings of NAACL 2006*, New York, NY, 2006.

[12] Tan, P.-N., Steinbach, M., and Kumar, V.,*Introduction to Data Mining*, Addison-Wesley,2005.

[13] Bijankhan, M., Sheykhzadegan, J., Bahrani, M., Ghayoomi, M., *Lessons from building a persian written corpus: Peykare*, Lang ResourEval. 45(2), 143–164, 2011.

[۱۴] مهدی محسنی، بهروز مینایی بیدگلی، **سیستم برچسب‌گذاری و ابهام زدایی خودکار اجزای کلام پیکره‌ی متنی زبان فارسی**، پایان‌نامه کارشناسی ارشد، دانشگاه علم و صنعت ایران، تهران، صفحه ۷۸، ۱۳۸۷.

زیرنویس‌ها

[1] Computational Linguistic
[2] Part of Speech Tagging
[3] Unknown Words
[4] Hidden Markov Model
[5] Association Rule
[6] Hidden Markov Model
[7] Natural Language Processing
[8] Unseen Words
[9] Out of Vocabulary Words
[10] Unknown Word Identification
[11] Unknown Word Syntactic Processing
[12] Unknown Word Meaning
[13] Unknown Word Sense Detection
[14] Heuristic
[15] Regular Expression
[16] Rule Set
[17] Support measure
[18] Confidence measure
[19] Interest
[20] Jaccard
[21] $\varphi - coefficent$
[22] Fuzzy
[23] Prior Probability