خوشه‌بندی خودکار کلمات بر اساس مقوله‌های نحوی برای سیستم‌های بازشناسی گفتار پیوسته فارسی

محمد بحرانی
آزمایشگاه پردازش گفتار
دانشکده مهندسی کامپیوتر دانشگاه صنعتی شریف
bahrani@ce.sharif.edu

حسین صامتی
استادیار و عضو هیات علمی
دانشکده مهندسی کامپیوتر
دانشگاه صنعتی شریف
sameti@sharif.edu

نازیلا حافظی
آزمایشگاه پردازش گفتار
دانشکده مهندسی کامپیوتر
دانشگاه صنعتی شریف
hafezi@ce.sharif.edu

سعیده ممتازی
آزمایشگاه پردازش گفتار
دانشکده مهندسی کامپیوتر دانشگاه صنعتی شریف
momtazi@ce.sharif.edu

چکیده: در این مقاله روش جدیدی برای خوشه‌بندی کلمات به منظور ساخت مدل زبانی n-gram برای زبان فارسی ارائه شده است که در آن مشکل پیچیدگی روش‌های خودکار و سر گشتگی بالای روش‌های دستی به حداقل رسیده است. در این روش هر کلمه با یک بردار ویژگی نمایش داده می‌شود که این بردار معرف آمار مقوله‌های نحوی مربوط به آن کلمه است. سپس بردارهای حاصل با استفاده از الگوریتم k-means خوشه‌بندی می‌شوند. پیاده سازی و آزمایش‌های مربوط بر روی پیکره متنی زبان فارسی که شامل حدود ۱۰میلیون کلمه می‌باشد، صورت گرفته است. نتایج بیانگر کاهش۳۴ درصدی در سرگشتگی و کاهش ۱۶ درصدی در نرخ خطای بازشناسی نسبت به روش‌های دستی مبتنی بر مقوله‌های نحوی است.

واژه‌های کلیدی: مدل n-gram خوشه‌بندی شده، بازشناسی گفتار پیوسته، مقوله نحوی، پیکره متنی زبان فارسی، خوشه‌بندی کلمات.

۱- مقدمه

یکی از مؤثرترین راه‌های افزایش دقت سیستم بازشناسی گفتار پیوسته، به‌کارگیری اطلاعات زبانی (به صورت آماری، گرامری و . . .) می‌باشد. روشهای مختلفی برای به‌کارگیری اطلاعات زبانی در بازشناسی گفتار وجود دارد. در این میان استفاده از مدل‌های زبانی آماری (به صورت n-gram) در سیستم‌های بازشناسی گفتار بسیار رایج بوده و نتایج خوبی داشته است [۱، ۲].

در مدل زبانی n-gram فرض بر این است که احتمال رخداد یک کلمه در متن فقط به n-1 کلمه قبلی آن وابسته است [۳]. برای استخراج این مدل برای هر زبان خاص نیاز به یک پیکره متنی[۱] شامل حجم عظیمی از داده‌های متنی آن زبان داریم. هر چه حجم داده‌های مورد استفاده بیشتر باشد تخمین بهتری از مدل‌های n-gram به دست می‌آید. چون همیشه میزان داده‌های در دسترس محدود می‌باشد معمولا n، برابر با ۱، ۲ و یا ۳ انتخاب می‌شود و مدل‌های آماری استخراج شده به ترتیب mono-gram، bi-gram و tri-gram نامیده می‌شوند.

علیرغم این که مدل‌های زبانی n-gram پاسخگوی بسیاری از اطلاعات مورد نیاز برای به دست آوردن مدل زبانی آماری هستند، ولی از آنجا که پیکره‌های متنی موجود برای زبان فارسی هنوز نو پا می‌باشند و دارای حجم نسبتا کمی از داده‌های متنی هستند. برای استخراج مدل‌های n-gram با کمبود داده مواجه هستیم که به تنکی[۲] زیاد مدل‌های n-gram منجر می‌شود. برای رفع این مشکل می‌توان از خوشه‌بندی کلمات استفاده کرد. به این صورت که اگر آمار bi-gram یا tri-gram کلمه‌ها کافی نبود می‌توانیم به آمار میان کلاس‌هایی که آن کلمات به آن‌ها تعلق دارند مراجعه کنیم. مدل‌های n-gram به دست آمده در این حالت، به نام "مدل n-gram خوشه‌بندی شده"[۳] شناخته می‌شوند [۷].

برای خوشه‌بندی کلمات، ساده ترین روش خوشه‌بندی آن‌ها بر اساس مقوله نحویشان می‌باشد. یعنی کلماتی که از نظر نحوی در یک مقوله (اسم، صفت، فعل و ...) جای دارند، در خوشه‌بندی نیز در یک خوشه قرار می‌گیرند. علاوه بر این روش، روش‌های خوشه‌بندی متنوع دیگری نیز وجود دارند که بر خلاف روش فوق کار خوشه‌بندی را به صورت خودکار انجام می‌دهند [۵، ۶، ۷]. در بسیاری از این روش‌ها عمل خوشه‌بندی بر اساس معیارهایی که از تئوری اطلاعات گرفته شده‌اند، صورت می‌گیرد.

در این مقاله روش جدیدی برای خوشه‌بندی خودکار کلمات به منظور ساخت مدل زبانی مناسب، برای سیستم‌های بازشناسی گفتار پیوسته، پیشنهاد شده است. این روش در واقع از دو روش ذکر شده، یعنی خوشه‌بندی بر مبنای مقوله نحوی و خوشه‌بندی خودکار ایده گرفته است و تلفیقی از این دو را به عنوان راهکار ارائه داده است.

دادگان متنی مورد استفاده در این پژوهش، "پیکره متنی زبان فارسی" [۱۳] می‌باشد. توضیح بیشتر راجع به این دادگان در بخش ۲ مطرح می‌شود. در بخش ۳ به اختصار روش‌های خوشه‌بندی کلمات عنوان می‌شود. بخش ۴ شامل توضیح ایده تلفیقی جدید و مزایای آن است. در بخش ۵ نتایج آزمایش‌ها و مقایسه آن با روش‌های موجود عنوان می‌شود و در نهایت نتیجه گیری در بخش ۶ ذکر خواهد شد.

دادگان متنی مورد استفاده در این پژوهش نسخه اولیه "پیکره متنی زبان فارسی" می‌باشد که در "مرکز تحقیقات پردازش هوشمند علائم" تهیه شده است. این "پیکره متنی شامل مجموعه‌ای از متون مختلف زبان فارسی، در حجمی در حدود ۱۰ میلیون کلمه به همراه برچسب‌دهی نحوی هر کلمه است. برچسب نحوی هر کلمه نشانگر مقوله نحوی (POS[4]) و در صورت لزوم زیر مقوله‌های نحوی و معنایی آن کلمه هستند. به طور کلی ۸۸۲ برچسب در پیکره متنی وجود دارد. علاوه بر این‌که این تعداد برچسب بسیار زیاد و بعضی از آن‌ها بیش از حد جزئی می‌باشند، تعداد قابل ملاحظه‌ای از آن‌ها نیز بسیار کم به کار رفته‌اند؛ بنابراین با بررسی برچسب‌ها و استخراج آمار هر برچسب، آن‌ها را بر اساس تشابهشان (از لحاظ نحوی) به ۱۶۶ دسته کلی تقسیم نمودیم و به هر دسته، یک برچسب کلی اختصاص دادیم [۱۴]. با این کار هم تعداد کل برچسب‌ها کاهش می‌یابد و هم برچسب‌های کم‌کاربرد در دسته‌های کلی‌تر جای می‌گیرند

۳- مدل n-gram خوشه‌بندی شده

همانطور که درمقدمه ذکر شد، به دلیل کمبود داده‌های متنی نیاز به استفاده از مدل‌های n-gram خوشه‌بندی شده داریم. در این مدل‌ها احتمال bi-gram دو کلمه w_1 و w_2 که به ترتیب متعلق به کلاس‌های c_1 و c_2 باشند، به صورت زیر محاسبه می‌شود [۲]:

$$P(w_n \mid w_1 w_2 \cdots w_{n-1}) = P(c_n \mid c_1 c_2 \cdots c_{n-1}).P(w_n \mid c_n) \qquad (۱)$$

برای خوشه‌بندی کلمات و تعیین کلاسی که هر کلمه به آن متعلق است دو روش کلی وجود دارد. یکی روش خودکار و دیگری روش دستی. در روش‌های خودکاری که تاکنون ارائه شده اند عمدتا خوشه‌بندی طوری انجام می‌گیرد که معیارهایی مثل سرگشتگی[5] یا متوسط اطلاعات متقابل بین کلمات بهینه شوند. مشهورترین این روش‌ها، الگوریتم‌های خوشه‌بندی براون [۵] و مارتین [۶] می‌باشند که بار محاسباتی بالای آن‌ها کار پیاده سازی را برای تعداد خوشه‌های زیاد، به خصوص هنگامی که اندازه واژگان بزرگ باشد، دشوار می‌سازد. علاوه بر این روش‌ها که بر روی تئوری اطلاعات تکیه دارند روش‌ها معدودی نیز به چشم می‌خورد که در آن‌ها کلمات به عنوان بردارهایی در نظر گرفته شده‌اند و دسته‌بندی بر مبنای میزان شباهت آن بردارها صورت می‌گیرد. معروف‌ترین این روش‌ها روش کورکماز [۷] است که در آن از احتمال bi-gram هر کلمه نسبت به سایر کلمات به عنوان بردار ویژگی آن کلمه استفاده شده است.

روش‌های دستی موجود برای خوشه‌بندی کلمات عمدتا از مقوله‌های نحوی و معنایی کلمات استفاده می‌کنند. روش خوشه‌بندی مبتنی بر مقوله نحوی که رایج‌ترین روش خوشه‌بندی دستی می‌باشد بسیار ساده بوده و با استفاده از یک دادگان متنی برچسب‌دهی شده (که در آن بر چسب‌ها نشان دهنده مقوله نحوی کلمات هستند)، انجام می‌شود [۴، ۱۲]. یکی از مهم‌ترین معایب این روش این است که یک کلمه در بنابراین در چند خوشه مختلف قرار گیرد. این عیب منجر به سرگشتگی بیشتر این روش نسبت به روش‌های خودکار می‌شود. جدول ۱ مقایسه مختصری روی روش‌های خودکار و روش‌های دستی نشان می‌دهد.

جدول (۱): مقایسه روش خوشه‌بندی دستی و خودکار

	خوشه‌بندی خودکار	خوشه‌بندی دستی
بار محاسباتی	خیلی زیاد	خیلی کم
سرگشتگی	پایین	بالا
تعلق هر کلمه به خوشه‌ها	تنها به یک خوشه	بیش از یک خوشه

جدول فوق حاکی از این است که بار محاسباتی بالا در روش‌های خودکار مبتنی بر تئوری اطلاعات و همین طور سرگشتگی بالا و تعلق به بیش از یک کلاس، در روش‌های خوشه‌بندی دستی مبتنی بر مقوله‌های نحوی، مهم‌ترین معایب این روش‌ها هستند.

با توجه به این مقایسه، این ایده مطرح می‌شود که می‌توان با ترکیب این دو روش به روش جدیدی برای خوشه‌بندی کلمات رسید و در آن از اصول خوشه‌بندی خودکار و همین طور مفهوم خوشه‌بندی بر مبنای مقوله نحوی استفاده کرد. ضمن اینکه این تلفیق باعث در کنار هم قرار گرفتن مزایا و حذف معایب هر یک از این دو روش گردد.

۴- روش پیشنهادی

همان‌طور که ذکر شد یک کلمه در بافت‌های متنی مختلف می‌تواند متعلق به مقوله‌های نحوی مختلفی باشد. با استفاده از "پیکره متنی زبان فارسی" می‌توان آمار مقوله‌های نحوی مختلف برای هر کلمه را استخراج کرد. بدین منظور به ترتیب زیر عمل شده است [۱۴]:

۱. انواع کلمات مختلف پیکره متنی همراه با تعداد رخداد آن‌ها استخراج شده‌اند.

۲. با توجه به تعداد رخداد هر کلمه، ۲۰۰۰۰ کلمه پرکاربرد مشخص شده و به عنوان مجموعه لغات در نظر گرفته شده است. سایر لغات همگی به عنوان یک لغت "خارج از واژگان" به حساب می‌آیند.

۳. برای ۲۰۰۰۰ کلمه موجود در مجموعه لغات، آمار رخداد هر برچسب (مقوله نحوی) برای آن کلمه استخراج شده است. با توجه به این که تعداد مقوله‌های نحوی را در پیکره متنی به ۱۶۶ مقوله کاهش داده‌ایم، آمار مذکور به صورت یک ماتریس ۱۶۶×۲۰۰۰۰ ذخیره می‌شود. این ماتریس را ماتریس POS نام نهاده‌ایم.

به عنوان مثال کلمه "زیبا" در پیکره متنی، ۹ بار با برچسب "قید غیر کمّی ساده"، ۷ بار با برچسب "اسم مفرد خاص" و ۴۲۰ بار با برچسب "صفت ساده" به کار رفته است. آمار سایر مقوله‌های نحوی برای این کلمه صفر می‌باشد. بنابراین آن سطر از ماتریس POS که مربوط به کلمه "زیبا" می‌باشد، در سه خانه دارای مقادیر ۷، ۹ و ۴۲۰ بوده و سایر خانه‌های آن صفر می‌باشد. همچنین کلمه "مرد" نیز ۱۲۰۱ بار به عنوان "اسم مفرد عام" ، ۳۲ بار به عنوان "صفت ساده" و ۴۷ بار به عنوان "فعل

کلمه مقوله‌های نحوی متفاوت بگیرد، هم نگاره بودن دو کلمه "مَرد" و "مُرد" در زبان فارسی می‌باشد. از آنجا که تفکیک آمار کلمات هم‌نگاره فارسی به طور خودکار حتی با در نظر گرفتن مقوله‌های نحوی آن‌ها کار ساده‌ای نیست، در مقاله حاضر از این کار صرف‌نظر شده است.

در این پژوهش قصد داریم از این ویژگی که هر کلمه می‌تواند چند مقوله نحوی مختلف بگیرد در راستایی استفاده کنیم که منجر به کاهش سرگشتگی شود. برای رسیدن به این هدف می‌توان از روش‌های خودکار مبتنی بر بردار ویژگی استفاده کرد. به این صورت که سطر مربوط به هر کلمه در ماتریس POS را به عنوان بردار ویژگی آن کلمه در نظر گرفته‌ایم و از این بردارهای ویژگی برای خوشه‌بندی کلمات استفاده کرده‌ایم. با توجه به مثال‌های فوق، دلیل استفاده از این بردار برای خوشه‌بندی کلمات مشخص می‌شود. منطقی به نظر می‌رسد که اگر دو کلمه دارای توزیع نحوی مشابه باشند باید در یک دسته جای گیرند و کلمات با توزیع نحوی مشابه دارای بردارهای ویژگی مشابه می‌باشند. به عنوان مثال کلماتی که همیشه به عنوان فعل ماضی ساده و یا اسم مفرد خاص به کار می‌روند در یک دسته جا داده می‌شوند.

برای خوشه‌بندی بردارها از الگوریتم خوشه‌بندی برداری k-means استفاده شده است. در این الگوریتم ابتدا یک خوشه‌بندی اولیه از بردارها ایجاد شده و مرکز خوشه‌ها تعیین می‌گردد. سپس خوشه‌بندی انجام شده اصلاح می‌شود. به این صورت که هر بردار به خوشه‌ای تعلق می‌گیرد که نزدیکترین فاصله را با مرکز آن خوشه داشته باشد. در مرحله بعد مرکز خوشه‌های جدید محاسبه شده و رویه مذکور تکرار می‌گردد تا زمانی که دیگر تغییری در خوشه‌ها صورت نگیرد.

برای تعیین فاصله بردارها در الگوریتم k-means از تابع فاصله اقلیدسی استفاده شده است. مرکز خوشه‌ها نیز با میانگین‌گیری معمولی از بردارهای موجود در آن‌ها به دست می‌آید. قابل ذکر است که در الگوریتم k-means باید تعداد خوشه‌ها را از پیش تعیین کرد. بنابراین در این پروژه روند خوشه‌بندی را با تعداد خوشه‌های از پیش تعیین شده مختلفی انجام داده‌ایم.

مزیت این روش سادگی الگوریتم از لحاظ پیچیدگی زمانی نسبت به سایر الگوریتم‌های رایج خوشه‌بندی کلمات (مانند الگوریتم مارتین و براون) و همچنین درنظرگرفتن مقوله‌های نحوی کلمات در انتساب آن‌ها به خوشه‌های مختلف می‌باشد. با این حال این روش تضمین می‌کند که هر کلمه فقط در یک خوشه جای گیرد. در روش‌های مشابهی که از نمایش برداری برای کلمات و الگوریتم‌های خوشه‌بندی برداری استفاده می‌کنند (مانند روش کورکماز) از احتمالات bi-gram هر کلمه به عنوان بردار ویژگی استفاده می‌شود. در این روش‌ها اگر اندازه مجموعه لغات بزرگ باشد ابعاد بردارها خیلی بزرگ شده و رویه دسته‌بندی بسیار کند می‌گردد در حالی که ابعاد بردارها در روش پیشنهادی کوچک بوده و بستگی به اندازه مجموعه لغات ندارد.

"پیکره متنی زبان فارسی" (حدود ۸،۵ میلیون کلمه) و باتوجه با کلاس مربوط به هر کلمه، مدل‌های n-gram خوشه‌بندی شده قابل استخراج می‌باشند. در این مقاله با توجه به حجم پیکره متنی به مدل‌های bi-gram و tri-gram خوشه‌بندی شده اکتفا شده است.

۵- نتایج به دست آمده

برای ارزیابی مدل‌های n-gram حاصل از خوشه‌بندی از دو روش استفاده شده است. اول میزان سرگشتگی حاصل از این مدل‌ها و دوم تاثیر نهایی آن‌ها در افزایش دقت بازشناسی گفتار. برای اندازه‌گیری میزان سرگشتگی، حدود ۴۳۸۰۰ جمله از پیکره متنی (در حدود یک میلیون کلمه) به عنوان مجموعهٔ تست انتخاب شد (لازم به ذکر است که از این جملات در هنگام استخراج آمار n-gram استفاده نشده است). سپس سرگشتگی حاصل از مدل‌ها با استفاده از رابطه کلی زیر محاسبه گردید:

$$PP = \hat{P}(w_1 w_2 w_3 \cdots w_m)^{-1/m} \qquad (۲)$$

در این رابطه، m تعداد کلمات مجموعه تست و رشته $w_1 w_2 w_3 \cdots w_m$ نشان‌دهنده دنباله کلمات این مجموعه می‌باشد. احتمال $\hat{P}(w_1 w_2 w_3 \cdots w_m)$ بر حسب درجه مدل آماری (bi-gram یا tri-gram) به صورت‌های زیر ساده می‌شود:

$$PP_{bi} = \left(\prod_{i=1}^{m} \hat{P}(w_i \mid w_{i-1}) \right)^{-1/m} \qquad (۳)$$

$$PP_{tri} = \left(\prod_{i=1}^{m} \hat{P}(w_i \mid w_{i-1} w_{i-2}) \right)^{-1/m} \qquad (۴)$$

که احتمال‌های $\hat{P}(w_i \mid w_{i-1})$ و $\hat{P}(w_i \mid w_{i-1} w_{i-2})$ از رابطه ۱ به دست می‌آیند.

جدول زیر سرگشتگی حاصل از این مدل‌ها را به ازای تعداد کلاس‌های مختلف نشان می‌دهد. برای مقایسه روش خوشه‌بندی پیشنهادی با روش دستی مبتنی بر مقوله‌های نحوی، سرگشتگی حاصل از مدل‌های n-gram به دست آمده از روش دستی نیز در جدول آمده است. همان طور که مشاهده می‌شود استفاده از این روش بهبود قابل ملاحظه ای را در سرگشتگی نتیجه داده است.

جدول (۲): سرگشتگی حاصل از مدل‌های n-gram مختلف بر روی ۴۳۸۰۰ جمله پیکره متنی

مدل آماری/تعداد کلاس‌ها	Bi-gram perplexity	Tri-gram perplexity
خوشه‌بندی‌شده (۱۰۰ کلاس)	۷۳۹	۶۸۱
خوشه‌بندی‌شده (۲۰۰ کلاس)	۶۹۹	۶۴۴
خوشه‌بندی‌شده (۵۰۰ کلاس)	۶۵۲	۵۸۹
مبتنی بر POS	۹۹۰	۹۱۱

برای آزمایش میزان تاثیر مدل‌های n-gram خوشه‌بندی شده بر دقت

استفاده شده است. این سیستم بازشناسی، مبتنی بر مدل مخفی مارکوف بوده و برای بازشناسی کلمات از الگوریتم "جستجوی همزمان واج و کلمه مبتنی بر درخت واژگان" [۱۰] استفاده می‌کند. دادگان گفتاری مورد استفاده برای آموزش مدل‌ها دادگان فارس‌دات [۱۱] می‌باشد. برای انجام آزمایش‌ها، ۱۴۰ جمله از این دادگان کنار گذاشته شد و آموزش مدل‌های مخفی مارکوف برای هر یک از ۳۰ واج زبان فارسی با استفاده از بقیه دادگان فارس‌دات (۵۹۴۰ جمله) انجام شد. اندازه مجموعه واژگان مورد استفاده ۱۰۹۲ کلمه (کلمات دادگان فارس‌دات) بوده و آزمایش‌ها بر روی ۱۴۰ جملهٔ کنار گذاشته شده انجام گرفته است.

مدل‌های n-gram خوشه‌بندی شده به روش "در حین جستجو" [۸] در سیستم بازشناسی به کار گرفته شده‌اند. بدین صورت که هنگامی که رویه جستجو فرضیه‌های مختلف را برای بازشناسی کلمات به پیش می‌برد، به ازای هر کلمه بازشناسی شده، امتیازهای حاصل از مدل زبانی برای آن کلمه با امتیاز حاصل از مدل آکوستیک ترکیب می‌شود و امتیاز جدید فرضیه را تشکیل می‌دهد.

جدول زیر دقت و صحت بازشناسی را در هنگام استفاده از مدل‌های n-gram حاصل از خوشه‌بندی به روش پیشنهادی در مقایسه با به‌کارگیری مدل‌های حاصل از خوشه‌بندی به روش دستی نشان می‌دهد. همان‌طور که مشاهده می‌شود تاثیر مدل‌های n-gram خوشه‌بندی شده بر دقت بازشناسی قابل توجه است.

جدول (۳): درصد میزان دقت و صحت به‌دست‌آمده با به‌کارگیری مدل‌های خوشه‌بندی مختلف در سیستم بازشناسی گفتار

نوع مدل زبانی	تعداد کلاس‌ها	دقت (٪)	صحت (٪)
مدل bi-gram خوشه‌بندی‌شده	۱۰۰	۷۵/۱۳	۸۱/۷۲
	۲۰۰	۷۵/۶۰	۸۲/۰۷
	۵۰۰	۷۶/۲۲	۸۲/۵۷
مدل tri-gram خوشه‌بندی‌شده	۱۰۰	۷۶/۵۵	۸۲/۰۲
	۲۰۰	۷۶/۶۷	۸۲/۱۴
	۵۰۰	۷۷/۰۲	۸۲/۹۱
مدل bi-gram مبتنی بر POS	۱۶۶	۷۲/۶۴	۷۹/۴۵

۶- نتیجه‌گیری

در این مقاله روش جدیدی برای خوشه‌بندی خودکار کلمات فارسی بر اساس مقوله‌های نحوی آن‌ها ارائه شد و مدل‌های n-gram حاصل از خوشه‌بندی در یک سیستم بازشناسی گفتار پیوسته مورد ارزیابی قرار گرفت. این روش از لحاظ پیچیدگی محاسباتی نسبت به سایر روش‌های خودکار کم هزینه بوده و برای پیاده‌سازی بسیار ساده می‌باشد. ارزیابی مدل‌های زبانی حاصل، نشان‌دهنده کاهش در سرگشتگی و همچنین کاهش در نرخ خطای بازشناسی نسبت به روش دستی مبتنی بر مقوله‌های نحوی می‌باشد.

[۱] محمود بیجن‌خان، **پیکره متنی زبان فارسی**، پژوهشکده پردازش هوشمند علائم، ۱۳۸۳.

[۲] محمد بحرانی، حسین ثامتی، نازیلا حافظی، سعیده ممتازی، حامد موثق، به‌کارگیری پیکره متنی زبان فارسی در ساخت مدل‌های زبانی آماری برای سیستم‌های بازشناسی گفتار پیوسته فارسی، **مجموعه مقالات دومین کارگاه پژوهشی زبان فارسی و رایانه**، صص ۹۲-۱۰۹، تهران، ۱۳۸۵.

[3] X. Huang, F. Alleva, H. Hon, M. Hwang, K. Lee and R. Rosenfield, The SPHINX-II Speech Recognition System: An Overview, *Computer Speech and Langauge*, vol. 2, pp. 137-148, 1993.

[4] S. J. Young, J. Jansen, J.J. Odell, D. Ollason, and P.C. Woodland, *The HTK Hidden Markov Model Toolkit Book*, 1995.

[5] L. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, New Jersey: Prentice Hall, 1993.

[6] P. A. Heeman, POS tagging versus Classes in Language Modeling, *Proc. 6th Workshop on Very Large Corpora*, pp. 179-187, Aug. 1998.

[7] P. Brown, V. Della Pietra, P. deSouza, J. Lai, and Robert L. Mercer, Class-based n-gram models of natural language, *Computational Linguistics* 18(4), pp. 467–479, 1992.

[8] S. Martin, J. Liermann, H. Ney, Algorithms for bigram and trigram word clustering, *Speech Communication*, vol. 24, pp. 19-37, 1998.

[9] E. E. Korkmaz, G. Ucoluk, A Method for Improving Automatic Word Categorization, *Workshop on Computational Natural Language Learning*, Madrid, Spain, pp. 43-49, 1997.

[10] M. P. Harper, L. H. Jamieson, C. D. Mitchell, G. Ying, Integrating Language Models with Speech Recognition, *AAAI-94 Workshop on the Integration of Natural Language and Speech Processing*, pp. 139-146, Aug. 1994.

[11] B. Babaali, H. Sameti, The Sharif Speaker-Independent Large Vocabulary Speech Recognition System, *The 2nd Workshop on Information Technology & Its Disciplines*, Kish Island, Iran, Feb. 24-26, 2004.

[12] H. Ney, R. Haeb-Umbach, B.-H. Tran, M. Oerder, Improvements in Beam Search for 10000-Word Continuous Speech Recognition, *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 13-16, March 1992.

[13] M. Bijankhan et al., FARSDAT-The Speech Database of Farsi Spoken Language, *In Proc. The 5th Australian Int. Conf. on Speech Science and Tech*., Vol. 2, Perth, 1994.

[14] M. Bahrani, H. Samet, N. Hafezi, H. Movasagh, Building and Incorporating Language Models for Persian Continuous Speech Recognition Systems, *In Proc. 5th international conference on Language Resources and Evaluati*, pp. 101-104, Genoa, Italy, May 2006.

زیرنویس‌ها

[1] text corpus
[2] sparseness
[3] class-based n-gram language model
[4] Part Of Speech
[5] perplexity