استخراج روابط محلی وابسته به ترتیب کلمات با استفاده از یک مدل سلسله‌مراتبی بیز

مرضیه رحیمی[۱]، دانشجوی دکتری؛ مرتضی زاهدی[۲]، استادیار؛ هدی مشایخی[۳]، استادیار

۱- دانشکده مهندسی کامپیوتر و فناوری اطلاعات - دانشگاه صنعتی شاهرود - شاهرود - ایران - marziea.rahimi@shahroodut.ac.ir
۲- دانشکده مهندسی کامپیوتر و فناوری اطلاعات - دانشگاه صنعتی شاهرود - شاهرود - ایران - zahedi@shahroodut.ac.ir
۳- دانشکده مهندسی کامپیوتر و فناوری اطلاعات - دانشگاه صنعتی شاهرود - شاهرود - ایران - hmashayekhi@shahroodut.ac.ir

چکیده: در این مقاله، یک مدل سلسله مراتبی بیز برای استخراج روابط محلی کلمات معرفی شده‌است. این مدل را می‌توان یک مدل برای زبان دانست. مدل‌های زبانی کنونی به دلیل وابستگی به ترتیب دقیق کلمات، به شدت از مشکل تنکی رنج می‌برند. مدل پیشنهادی قادر است ضمن نادیده نگرفتن ترتیب کلمات، این مشکل را تخفیف دهد. در مدل پیشنهادی که یک مدل مولد است، فرض می‌شود که هر کلمه از یکی از کلمات قبلی خود در یک بازه محدود یا به‌بیان‌دیگر، یک پنجره با طول ثابت، تولید شده‌است. به‌این‌ترتیب، هر کلمه خود توزیعی بر روی کلمات است. برخلاف مدل‌های n-gram که توزیعی بر روی دنباله‌های کلمات هستند و درنتیجه دنباله‌های دقیقاً مرتب کلمات را می‌شمرند، در مدل پیشنهادی به دنبال زوج‌کلماتی هستیم که ممکن است با فاصله‌های مختلف از یکدیگر رخ داده باشند. به‌این‌ترتیب مشکل تنکی تا حد زیادی تخفیف می‌یابد. مدل پیشنهادی از نظر تواناییش در مدل کردن داده‌ها با استفاده از معیار perplexity با مدل n-gram مقایسه شده‌است و برای پنجره‌هایی با طول‌های مختلف، بهتر از مدل n-gram عمل کرده‌است.

واژه‌های کلیدی: مدل‌های سلسله مراتبی بیز، مدل‌های گرافیکی، نمونه‌برداری گیبس، مدل‌های زبانی، زنجیره مارکوف مونت کارلو، روابط کلمات.

Extracting Order-Sensitive Word-to-Word Relations Using a Hierarchical Bayes Model

Marziea Rahimi[1], PhD Student; Morteza Zahedi[2], Assistant Professor; Hoda Mashayekhi[3], Assistant Professor

1- Department of Computer and IT Engineering, Shahrood University of Technology, Shahrood, Iran, Email: marziea.rahimi@shahroodut.ac.ir
2- Department of Computer and IT Engineering, Shahrood University of Technology, Shahrood, Iran, Email: zahedi@shahroodut.ac.ir
3- Department of Computer and IT Engineering, Shahrood University of Technology, Shahrood, Iran, Email: hmashayekhi@shahroodut.ac.ir

Abstract: In this paper, a hierarchical Bayes model is introduced which models local word relationships in a language. The model can be considered as a language model. The proposed model does not suffer from sparseness because it does not rely on the exact word orders. However, the model does not completely ignore the word orders. The proposed generative model assumes that each word is a distribution over words and the current word is generated from the distribution of one of its previous words located in a fixed-size window. Contrary to an n-gram model which is a distribution over word sequences and so takes the exact sequences of words into account, the proposed model considers ordered pairs of words which can occur at different distances in the subject text data. Because of this, the sparseness problem is not severe for the proposed model. The model is compared with and outperformed n-gram model according to its ability to model text data which is evaluated by perplexity.

Keywords: Hierarchical Bayes models, Gibbs sampling, graphical models, language models, Markov chain Monte Carlo, word-to-word relationships.

تاریخ ارسال مقاله:
تاریخ اصلاح مقاله:
تاریخ پذیرش مقاله:
نام نویسنده مسئول:
نشانی نویسنده مسئول: ایران – سمنان- شاهرود - بلوار دانشگاه - دانشگاه صنعتی شاهرود- پردیس ۱- دانشکده مهندسی کامپیوتر و فناوری اطلاعات – آزمایشگاه وب و داده‌کاوی.

۱- مقدمه

با همه‌گیر شدن اینترنت و با توجه به اینکه بیشتر داده‌های موجود در اینترنت به‌صورت متن هستند، حجم عظیمی از داده‌های متنی الکترونیک در اختیار بشر قرار گرفته که روزبروز در حال افزایش هستند. وجود این حجم عظیم داده امکان انجام تحلیل‌ها و استخراج اطلاعاتی را فراهم کرده‌است که پیش از این ممکن نبوده‌اند. توسعه مدل‌های احتمالاتی زبانی یکی از این امکان‌های فراهم شده‌است.

مدل‌های احتمالاتی زبانی یا همان مدل‌های زبانی، مکانیزم‌های احتمالاتی برای تولید متن در یک زبان هستند [۱]. براساس این تعریف، گستره وسیعی از مدل‌های زبانی قابل‌تصور است. موردتوجه‌ترین و پرکاربردترین مدل‌های زبانی معرفی‌شده تاکنون مدل‌های n-gram هستند و توسعه مدل‌های زبانی عموماً در جهت بهبود همین مدل‌ها بوده است. خواستگاه مدل‌های n-gram، حوزه شناسایی صحبت است که در آن سعی می‌شود دنباله‌های اصوات با دنباله‌های کلمات نگاشت داده شوند. بنابراین مدل‌های n-gram مبتنی بر ترتیب دقیق کلمات هستند. این مدل‌ها در واقع سعی می‌کنند با توجه به یک دنباله nتایی از کلمات یک عبارت، کلمه بعدی را پیش‌بینی کنند. از آنجا که این مدل‌ها ترتیب دقیق کلمات را در نظر می‌گیرند، مشکل تنکی برای آنها یک مشکل جدی محسوب می‌شود، به‌ویژه برای مقادیر بزرگ n. این مشکل از آنجا ناشی می‌شود که بسیاری از ترکیبات ممکن کلمات در مجموعه داده رخ نخواهند داد و بنابراین برای ساخت یک مدل قابل اعتماد احتیاج به حجم بزرگی از داده است.

در کاربردهایی مانند شناسایی صحبت و ترجمه ماشینی، ترتیب دقیق کلمات دارای نقش کلیدی است و بنابراین مدل‌های n-gram کاملاً مناسب به‌کارگیری در چنین حوزه‌هایی هستند. در دهه اخیر، این مدل‌ها در کاربردهای دیگری مثل بازیابی اطلاعات [۱] و استخراج مدل‌های موضوعی [۲] نیز به کار گرفته و مؤثر نیز واقع شده‌اند. در این کاربردها، ترتیب کلمات در برخی موارد مثلاً در مورد اصطلاحات یا ترکیبات خاص می‌تواند کارآمد باشد. بااین‌حال، ترتیب دقیق کلمات، در چنین کاربردهایی نقش مستقیم و حیاتی بازی نمی‌کند [۳]. باتوجه به اینکه استفاده از مدل‌های n-gram ما را با مشکل تنکی مواجه خواهد نمود، مهم است بتوانیم مدلی از زبان بسازیم که ضمن نادیده نگرفتن ترتیب کلمات بتواند مشکل تنکی را تخفیف دهد. در این مقاله، هدف ما معرفی چنین مدلی است. مدل پیشنهادی فرض می‌کند که هر کلمه بر مبنای یکی از کلمات پیش از خودش تولید شده‌است. این کلمه پیشین می‌تواند در هر موقعیتی، در یک محدوده خاص قبل از کلمه جاری، قرار گرفته باشد. به‌همین‌دلیل، تغییر ترتیب کلمات متن، کلمات پیشین هر کلمه را تغییر خواهد داد و بنابراین مدل، مستقل از ترتیب کلمات متن نیست. بااین‌حال، از آنجا که دنباله‌های دقیقاً مرتب کلمات را در نظر نمی‌گیریم، مشکل تنکی تا حد زیادی تخفیف می‌یابد. در مدل پیشنهادی هر کلمه خود توزیعی بر روی کلمات است و کلمه جاری از توزیع مربوط به یکی از کلمات قبلیش تولید می‌شود. برای ارائه چنین مدلی از مفهوم مدل‌های سلسله‌مراتبی بیز استفاده شده‌است. یکی از مشوق‌های ما برای استفاده از این نوع مدل‌ها این است که زبان‌های طبیعی ماهیتاً دارای ساختار سلسله‌مراتبی هستند [۴، ۵]. زبان به عنوان یک ساختار سلسله‌مراتبی از چند منظر قابل توصیف است. یک منظر، ساختار سلسله‌مراتبی زبان را در واحدهای سازنده آن می‌بینند. به عنوان مثال واژها به هم پیوسته و کلمات را می‌سازند، کلمات در کنار یکدیگر عبارات را ساخته و جملات از به هم پیوستن عبارات ساخته می‌شوند. این منظر که بیشتر مناسب تشخیص گفتار است الهام‌بخش تلاش‌هایی [۴] در جهت به‌کارگیری مدل‌های احتمالاتی سلسله‌مراتبی در کاربرد مذکور شده‌است.

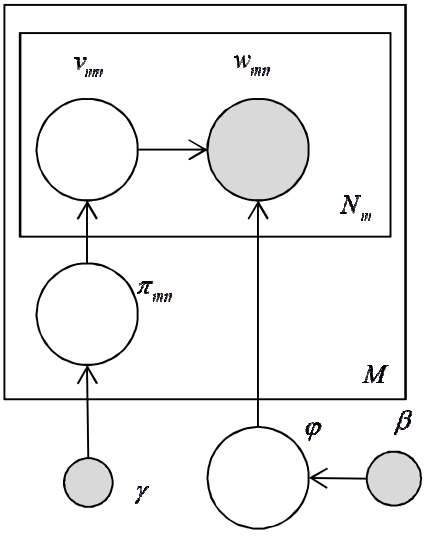
منظر دیگر ساختار سلسله‌مراتبی زبان را در خوشه‌های معنادار واحدهای زبانی می‌بینند. این نوع نگاه در کنار مفهوم مدل‌های سلسله‌مراتبی احتمالاتی منجر به شکل‌گیری مدل‌های موضوعی احتمالاتی [۶] شده و همچنین الهام‌بخش مدل پیشنهادی این مقاله است.

مدل‌های موضوعی احتمالاتی فرض می‌کنند که برای هر مجموعه متن، تعدادی موضوع داریم که هر یک توزیعی بر روی کلمات است. به هر کلمه از هر سند یک موضوع اختصاص می‌یابد و کلمه مربوطه از توزیع مربوط به آن موضوع استخراج می‌گردد. این مدل‌ها در واقع مدل‌هایی مولد برای متون یک زبان هستند که بر مبنای مدل‌های سلسله‌مراتبی بیز شکل گرفته‌اند. اولین مدل از این دسته مدل‌ها، مدل انتساب پنهان دریکله (LDA) [۶] است.

در مدل مذکور، ترتیب کلمات به طور کلی نادیده گرفته شده و اسناد مجموعه داده هدف، کیسه‌هایی از کلمات فرض می‌شوند. پس از معرفی LDA یکی از مسیرهایی که محققان در زمینه مدل‌های موضوعی در پیش گرفته‌اند، تلاش برای معرفی مدل‌هایی است که ترتیب کلمات در آنها نادیده گرفته نشود. بیشتر مدل‌های معرفی‌شده [۲، ۷] در این راستا مبتنی بر گنجاندن مدل‌های زبانی n-gram در کنار مفهوم موضوع است. اشکال این روش‌ها همانند مدل‌های n-gram، تنکی و همچنین در برخی موارد پیچیدگی محاسباتی است [۷].

چنانچه پیش از این ذکر شد، در مدل‌های موضوعی احتمالاتی، ترتیب کلمات به طور کامل نادیده گرفته می‌شود و روابط محلی در آنها ارجحیتی به روابط دورتر ندارند. مدل پیشنهادی این مقاله، مدلی از زبان است که ترتیب کلمات را به طور کلی نادیده نمی‌گیرد. تفاوت این مدل با یک مدل موضوعی مانند LDA این است که در این مدل فرض می‌شود که هر کلمه توسط یکی از کلمات پیشین خود تولید شده‌است و نه موضوع منتسب به آن. به‌این‌ترتیب در مدل پیشنهادی، هر کلمه خود توزیعی بر روی کلمات است. این مدل ضمن اینکه می‌تواند در کاربردهایی که ترتیب دقیق کلمات در آنها نقش کلیدی بازی نمی‌کند مفید باشد، همچنین به راحتی قابل‌گنجاندن در یک مدل موضوعی احتمالاتی مانند LDA خواهد بود.

در ادامه این مقاله، در بخش ۲ به معرفی پیشینه مدل‌های زبانی پرداخته‌ایم. سپس، در بخش ۳ مدل پیشنهادی معرفی و روش تخمین پارامترهای مدل توصیف شده‌است. در مقاله حاضر، مدل پیشنهادی را با مدل زبانی n-gram مقایسه می‌نماییم و بررسی کاربردهای آن را به مراحل بعدی کار وامی‌گذاریم. نتایج آزمایشات و مقایسه مدل پیشنهادی با مدل‌های n-gram بر روی چند مجموعه داده مختلف در بخش ۴ گزارش شده‌اند.



شکل ۱- نمایش گرافیکی مدل پیشنهادی

۲- بررسی کارهای پیشین

مدل‌های زبانی ابتدا از حوزه شناسایی صحبت به حوزه تحلیل متن وارد شدند. پرکاربردترین این مدل‌ها از ابتدا تا کنون مدل‌های n-gram بوده‌اند که نقشی اساسی در کاربردهایی از جمله شناسایی صحبت [۸]، نویسه‌خوانی نوری [۹]، اصلاح املای کلمات [۱۰] و ترجمه ماشینی دارند. یک مسئله اصلی که این مدل‌ها با آن روبرو هستند مسئله هموارسازی است. به دلیل اینکه مدل‌های n-gram مبتنی بر ترتیب دقیق کلمات هستند بسیاری از ترکیبات ممکن، در مجموعه داده مشاهده نخواهند شد و بنابراین نیازمندیم که روش‌های هموارسازی را تعیین نماییم که احتمال غیرصفر به چنین ترکیباتی اختصاص دهند. برخی از مهمترین روش‌های هموارسازی که تا کنون معرفی‌شده‌اند عبارتند از روش Knesser-Ney [۱۱]، روش absolute dscounting [۱۲]، روش Jelinek-Mercer [۱۳] و روش Knesser-Ney بهبودیافته [۱۴]. برای اطلاعات بیشتر و توصیف این روش‌ها می‌توان به [۱۴] مراجعه نمود.

سایر مدل‌های زبانی معرفی‌شده تاکنون نیز تا جایی که ما می‌دانیم برمبنای همین مدل‌های n-gram شکل گرفته‌اند ولی با شیوه‌های محاسباتی متفاوت مانند مدل‌های مبتنی بر شبکه‌های عصبی [۱۵] یا مدل کلمات پنهان [۱۶] که از متغیرهای پنهان و مدل‌های گرافیکی برای محاسبه مدل زبانی استفاده نموده‌است یا [۱۷] که از روش تخمین تغییراتی برای محاسبه مدل زبانی استفاده نموده‌است. به‌طورکلی، مدل‌های زبانی موجود را می‌توان به دو دسته تقسیم نمود: مدل‌های بلنددامنه و مدل‌های کوتاه‌دامنه. مدل‌های کوتاه‌دامنه مدل‌هایی هستند که روابط نزدیک کلمات را مدل می‌کنند یا همان مدل‌های n-gram با nهای کوچک. مدل‌های بلنددامنه مدل‌هایی هستند که مبتنی بر روابط دورترند، مثل مدل‌هایی که روابط ساختاری را درنظر می‌گیرند [۱۸] یا مدل‌های n-gram با nهای بزرگ. به‌طورکلی مدل‌های بلنددامنه دقیق‌تر هستند ولی به دلیل مشکل تنکی یا سنگینی محاسبات عموماً عملی نیستند [۱۷، ۱۹]. اگر مدل‌های زبانی را مکانیزم‌هایی احتمالاتی برای تولید متن در یک زبان بدانیم روش‌های گوناگونی را با زوایای دید متفاوت با مدل‌های n-gram می‌توان متصور شد. هرچند یکی از ویژگیهای مهم متن ترتیب کلمات آن است، ترتیب دقیق کلمات لزوماً در همه کاربردهای مرتبط با متون طبیعی، نقش مستقیم بازی نمی‌کند. بنابراین یکی از مدل‌هایی که می‌توان برای زبان ارائه کرد مدلی است که مبتنی بر دنباله دقیقاً مرتب کلمات نباشد. در مدل معرفی‌شده در این مقاله، به جای اینکه کلمه جاری در یک متن را وابسته به دنباله مرتب کلمات قبلیش ببینیم، آن را وابسته به یکی از کلمات قبلی فرض می‌کنیم. با ارائه چنین مدلی، می‌توان مشکل تنکی را تا حد زیادی تخفیف داد. ضمن اینکه با در نظر گرفتن کلمات قبل، ترتیب را هم به طور کامل نادیده نگرفته‌ایم.

۳- مدل پیشنهادی

چنان‌که پیش از این ذکر شد، در این مقاله مدلی معرفی می‌شود که فرض می‌کند هر کلمه خود توزیعی بر روی کلمات است. هر کلمه می‌تواند، از توزیع مربوط به هرکدام از کلمات پیشین خود، در یک پنجره با طول ازپیش‌تعیین‌شده، تولید شده‌باشد. هنگامی که انسانی در حال نگارش یک متن است، گذشته از دستور زبان، عوامل گوناگونی ممکن است او را در انتخاب کلمه بعدی متن هدایت نمایند. حداقل دوتا از این عوامل در تحلیل زبانهای طبیعی مورد توجه قرار گرفته‌اند: مفهوم کلی که برای نگارش متن در ذهن اوست یا همان موضوع متن و سایر کلماتی که پیش از این در متن استفاده کرده‌است. اولی منجر به پیشنهاد مدلهای موضوعی و دومی منجر به پیشنهاد مدل n-gram شده‌است. در این مقاله، قصد ما پرداختن به موضوع متن نیست بلکه تکیه ما مستقیماً بر رابطه کلمات است. بااین‌حال، ایده این مقاله، چنان‌که در مقدمه ذکرشد، با مدل n-gram متفاوت است. می‌خواهیم $p(w_i \mid w_j)$ را که احتمال رخداد کلمه w_i است بدست آوریم، به شرط اینکه w_j در یک محدوده مشخص قبل از موقعیت آن رخ داده‌باشد. به‌این‌ترتیب، در مدل پیشنهادی، هر کلمه خود توزیعی بر روی کلمات است. در مدل‌های n-gram فرض بر این است که یک کلمه از توزیعی بر روی دنباله مرتب کلمات قبلیش تولید می‌شود. در مدل پیشنهادی فرض بر این است که هر کلمه از یکی از کلمات قبلیش گرفته شده‌است ولی از آنجا که فقط به کلمات قبلی نگاه می‌کنیم و موقعیت کلمات دارای اهمیت است می‌توان گفت که ترتیب به کلی نادیده گرفته نشده‌است بلکه اثر آن تاحدزیادی خفیف شده‌است. در این مدل، هر سند توزیعی بسیار تنک بر روی کلمات است.

در مدل پیشنهادی، هدف کاهش ابعاد داده نیست، بلکه یافتن ارتباط کلمات به‌گونه‌ای است که ترتیب به‌صورتی خفیف در نظر گرفته شده‌باشد، درعین‌حال وابسته به ترتیب دقیق کلمات نباشیم. از طرفی، این فقط یک هم‌رخدادی ساده نیست بلکه ارتباط کلمات در درجات مختلف هم‌رخدادی به‌گونه‌ای بدست می‌آید که بر مبنای آن، قادریم هر کلمه را به‌صورت توزیعی بر روی کلماتی نمایش دهیم که در کل مجموعه اسناد در اطراف آن ظاهر شده‌اند. برای توصیف مدل پیشنهادی فرض می‌کنیم که مجموعه داده D متشکل از M سند d_m است. هرسند d_m متشکل از N_m کلمه w_{mn} است که در آن n، موقعیت کلمه در سند را نشان می‌دهد. مجموعه کلمات یکه[1] متن را با $V=\{v_1,v_2,...,v_{|V|}\}$ نشان می‌دهیم. هر کلمه w_{mn} منتسب به یک کلمه v_{mn} است که در پنجره‌ای به طول L، قبل از کلمه مربوطه واقع شده‌است. گراف مربوط به مدل پیشنهادی در شکل ۱ نمایش داده شده‌است. در این شکل، φ حاوی $|V|$ توزیع چندجمله‌ای برروی کلمات است که هریک متناظر با یکی از کلمات یکه مجموعه داده می‌باشد. پارامترهای این توزیع‌ها از یک توزیع دریکله با پارامتر β پیروی می‌کنند. همچنین در شکل مذکور π_m حاوی یک توزیع چندجمله‌ای بر روی کلمات برای سند d_m است که پارامترهای آن نیز از یک توزیع دریکله با پارامتر γ پیروی می‌کنند. فرایند مولد مدل پیشنهادی به شرح زیر است:

- برای هر سند d_m در مجموعه داده D:
  - ابتدا $\pi_m \sim Dirichlet(\gamma)$ انتخاب کن.
  - برای هر کلمه w_{mn} در این سند که برای آن، n نماینده موقعیتش در سند d_m است:
    - یکی از کلمات پیشینش در پنجره‌ای به طول L مانند $v_{mn} \sim Multinomial(\pi_m)$ را به عنوان کلمه مولد انتخاب کن.
    - کلمه جاری w_{mn} را از توزیع مربوط به v_{mn} یعنی $\varphi_{v_{mn}}$ انتخاب کن.

فرایند مولد فوق چگونگی تولید یک سند را از نگاه مدل پیشنهادی، با فرض دانستن پارامترهای مدل، نشان می‌دهد. همان‌طور که در این فرایند می‌بینیم، برای تولید هر کلمه از متن، یکی از کلمات قبلیش به طور تصادفی انتخاب می‌شود و کلمه هدف از توزیع این کلمه استخراج می‌گردد. برای محاسبه پارامترهای مدل، فرایند مذکور باید معکوس گردد یعنی پارامترهای مدل را به‌گونه‌ای می‌یابیم که احتمال اسناد برمبنای آن‌ها حداکثر شود یا به‌بیان‌دیگر در فرایند تولید هر یک از کلمات متن، منجر به انتخاب بهترین کلمه از بین کلمات قبلیش گردد.

۳-۱- تخمین پارامترهای مدل

همان‌طور که پیش از این گفته شد، مدل پیشنهادی یک مدل مولد برای اسناد مجموعه داده یا به عبارتی کلمات آن سند است. براین‌اساس باید پارامترهای مدل را به‌گونه‌ای بیابیم که احتمال رخداد این مجموعه را حداکثر نماید. احتمال رخداد یک سند بر مبنای مدل پیشنهادی، با توجه به مدل گرافیکی شکل ۱، به‌صورتی که در رابطه (۱) آمده است محاسبه می‌گردد.

$$p(D\mid\varphi)=\prod_{m=1}^{M}\int p(\pi_m)\Big(\prod_{n=1}^{N_m}\sum_{v_{mn}}p(v_{mn}\mid\pi_{mn})\,p\big(w_{mn}\mid v_{mn},\varphi_{v_{mn},w_{mn}}\big)\Big)d\pi_m \quad (۱)$$

انتگرال موجود در رابطه فوق به‌صورت مستقیم قابل‌محاسبه نیست بنابراین باید از روش‌های تخمینی برای محاسبه پارامترها استفاده نماییم. برای تخمین پارامترها در مدل‌های مشابه از روش‌های مختلفی [۶، ۲۰] استفاده شده‌است که بهترین عملکرد را روش نمونه‌برداری گیبس داشته است [۲۱]. زمانی که تولید مستقیم نمونه‌های یک توزیع احتمال دشوار یا غیرممکن باشد، از نمونه‌برداری گیبس که یک الگوریتم زنجیره مارکوف مونت کارلو است، برای تخمین نمونه مشاهدات [۲۲، ۲۳] استفاده می‌شود. این دنباله از مشاهدات می‌توانند در تخمین توزیع‌های احتمال متغیرهای پنهان یا برخی پارامترهای مدل به‌کار گرفته‌شوند.

نمونه‌برداری گیبس ابزاری برای استنباط‌های آماری به خصوص استنباط بیزی است. این الگوریتم یک زنجیره مارکوف از نمونه‌ها می‌سازد که در آن، هر حالت، یک نمونه از تمامی متغیرهای موجود است. در هر حالت، هر متغیر به شرط مقادیر سایر متغیرها به‌روز می‌شود و زمانی از یک حالت به حالت دیگر می‌رویم که مقادیر تمامی متغیرها به‌روز شده‌باشد. در این روش، هر نمونه به نمونه‌های نزدیکش وابسته است بنابراین معمولاً از تمام نمونه‌ها استفاده نمی‌شود بلکه مثلاً از هر ۱۰۰ نمونه یکی برداشته می‌شود. همچنین تعدادی از اولین نمونه‌های انتخاب‌شده دور ریخته می‌شوند. برای اینکه نمونه‌برداری گیبس را اعمال نماییم باید $p\left(v_{xy}\mid v_{-xy},w\right)$ را محاسبه کنیم که در آن x اندیس سند جاری و y اندیس کلمه جاری است. نماد $-xy$ نیز به معنی همه موقعیت‌ها غیر از موقعیت جاری یعنی xy است. این احتمال در الگوریتم شکل ۲ برای به روز کردن نمونه‌ها استفاده می‌شود.

۱- ورودی‌های مدل: اندازه پنجره L، مقادیر فراپارامترهای β و γ، حداکثر تعداد تکرار $maxIter$ و خروجی آن
۲- به‌صورت تصادفی هر کلمه متن را به یکی از کلمات قبلیش که در محدوده پنجره‌ای به طول L واقع شده‌است، منتسب کن.
۳- مقادیر اولیه شمارنده $n_v^{d_m}$ را که نماینده تعداد کلماتی است که در سند d_m به کلمه v انتساب یافته‌اند، مشخص کن.
۴- مقادیر اولیه شمارنده $n_{\cdot}^{d_m}$ را که نماینده تعداد کلماتی در سند d_m است، مشخص کن.
۵- مقادیر اولیه شمارنده n_w^v را که نماینده تعداد کلماتی مانند w است که در سرتاسر مجموعه داده به کلمه v انتساب یافته‌اند، مشخص کن.
۶- مقادیر اولیه شمارنده $n_{\cdot}^v$ را که نماینده تعداد کل کلماتی است که سرتاسر مجموعه داده به کلمه v انتساب یافته‌اند، مشخص کن.
۷- برای تعداد تکرار ۱ تا $maxIter$
۸- برای هر سند d_m
۹- برای هر موقعیت n در سند d_m
۱۰- آمار موقعیت جاری رااز شمارنده‌ها حذف کن.
۱۱- کلمات موجود در پنجره‌ای در موقعیت $n-1$ تا $n-L$ را مشخص کن.
۱۲- یکی از آنها را با توجه به توزیع $p(v_{mn} \mid w, v_{-mn})$ در رابطه ۶ انتخاب کن.
۱۳- کلمه جدید را به موقعیت جاری (یا همان کلمه w_{mn}) منسوب کن.
۱۴- شمارنده‌ها را با توجه به کلمه انتسابی جدید به‌روز کن.
۱۵- بعد از پایان تکرارها مقادیر پارامترهای مدل یعنی π و φ را براساس روابط ۷ و ۸ محاسبه کن.

شکل ۲- الگوریتم نمونه‌برداری گیبس مربوط به محاسبه پارامترهای مدل

احتمال $p\left(v_{xy} \mid v_{-xy}, w\right)$ به فرم زیر قابل محاسبه است که در آن v_{xy} کلمه منتسب به کلمه جاری در موقعیت x از سند d_y است و v_{-xy} نماینده کلمات منتسب به سایر کلمات غیر از کلمه جاری است.

$$p\left(v_{xy} \mid v_{-xy}, w\right) = \frac{p\left(v_{xy}, v_{-xy}, w\right)}{p\left(v_{-xy}, w\right)} \propto p(v, w) \qquad (۲)$$

همان‌طور که در رابطه فوق مشاهده می‌نمایید، مقدار مخرج برای مقادیر مختلف v_{xy} ثابت خواهد بود و بنابراین قابل‌حذف است. با توجه به مدل گرافیکی شکل ۱، احتمال صورت را می‌توان به‌صورت زیر شکست:

$$p(v, w) = \int p(\pi)\, p(v \mid \pi)\, d\pi \times \int p(\varphi)\, p(w \mid v, \varphi)\, d\varphi \qquad (۳)$$

انتگرال‌های فوق با توجه به اینکه توزیع دریکله مزدوج توزیع چندجمله‌ای است، به‌صورت زیر قابل محاسبه هستند:

$$\int p(\pi)\, p(v \mid \pi)\, d\pi = \left(\frac{\Gamma\left(|V|\gamma\right)}{\Gamma(\gamma)^{|V|}}\right)^M \prod_{m=1}^{M} \frac{\prod_{v=1}^{|V|} \Gamma\left(n_v^{d_m} + \gamma\right)}{\Gamma\left(\sum_{v=1}^{|V|} n_v^{d_m} + \gamma\right)} \qquad (۴)$$

$$\int p(\varphi)\, p(w \mid v, \varphi)\, d\varphi = \frac{\Gamma\left(|V|\beta\right)}{\Gamma(\beta)^{|V|}} \prod_{v_1=1}^{|V|} \frac{\prod_{v_2=1}^{|V|} \Gamma\left(n_{v_2}^{v_1} + \beta\right)}{\Gamma\left(\sum_{v_2=1}^{|V|} n_{v_2}^{v_1} + \beta\right)} \qquad (۵)$$

در روابط فوق، $n_v^{d_m}$ نماینده تعداد کلماتی در سند d_m است که کلمه v به آنها انتساب یافته است و $n_{v_2}^{v_1}$ تعداد کلمات v_2 در کل مجموعه داده است که کلمه v_1 به آنها انتساب یافته است. نماد $\Gamma(.)$ نماینده تابع استاندارد گاما است.

پس از ساده کردن رابطه‌های (۴) و (۵)، به‌ترتیب به کسرهای $\frac{n_{-xy,v_{xy}}^{d_x} + \gamma}{n_{-xy,\cdot}^{d_x} + |V|\gamma}$ و $\frac{n_{-xy,w_{xy}}^{v_{xy}} + \beta}{n_{-xy,\cdot}^{v_{xy}} + |V|\beta}$ می‌رسیم. یعنی:

$$p\left(v_{xy} \mid v_{-xy}, w\right) \propto \frac{n_{-xy,v_{xy}}^{d_x} + \gamma}{n_{-xy,\cdot}^{d_x} + |V|\gamma} \times \frac{n_{-xy,w_{xy}}^{v_{xy}} + \beta}{n_{-xy,\cdot}^{v_{xy}} + |V|\beta} \qquad (۶)$$

در روابط فوق $n_{-xy,v_{x,y}}^{d_x}$ نماینده تعداد دفعاتی است که کلمه‌ای در سند d_x به کلمه $v_{x,y}$ اختصاص یافته‌است، بدون احتساب کلمه موجود در موقعیت جاری ($-xy$). همچنین $n_{-xy,w_{xy}}^{v_{x,y}}$ نماینده تعدد دفعاتی است که کلمه $w_{x,y}$ به کلمه $v_{x,y}$ از بین کلمات قبل از خودش، اختصاص یافته است بدون در نظر گرفتن کلمه جاری و $n_{-xy,\cdot}^{v_{x,y}}$ تعداد دفعاتی است کلمه‌ای در سرتاسر مجموعه داده به موضوع $v_{x,y}$ اختصاص یافته‌است، بازهم بدون احتساب موقعیت جاری. پارامترهای مدل برای هر یک از نمونه‌ها، با استفاده از روابط (۷) و (۸) قابل‌محاسبه هستند.

$$\pi_{mv} = \frac{n_v^{d_m} + \gamma}{n_\cdot^{d_m} + |V|\gamma} \quad (۷)$$

$$\varphi_{v_1,v_2} = \frac{n_{v_2}^{v_1} + \beta}{n_\cdot^{v_1} + |V|\beta} \quad (۸)$$

براساس روابط بدست‌آمده و توضیحات داده‌شده، در نهایت، می‌توان الگوریتم نمونه‌برداری گیبس برای مدل را به‌صورتی که در شکل ۲ آمده است، توصیف کرد. چنان‌که در این شکل قابل‌مشاهده است، پنجره‌ای که در طول توصیف مدل ذکر شد در نمایش گرافیکی ظهور نمی‌یابد ولی در طول محاسبات، فضای حالات را محدود می‌کند. در روابط فوق، β و γ پارامترهای توزیع‌های دریکله یا همان فراپارامترهای مدل هستند. در این مقاله، چنان‌که در مدل‌های مشابه مرسوم است، این پارامترها ثابت و متقارن در نظر گرفته شده‌اند.

در مدل پیشنهادی می‌خواهیم که تعداد اندکی از کلمات متن در تولید هر سند مؤثر باشند و همچنین می‌خواهیم که ارتباط هر کلمه با تعداد اندکی از کلمات دیگر مجموعه قوی و با بیشتر آنها ضعیف باشد. بنابراین می‌خواهیم که ماتریسهای φ و θ هردو تنک باشند. این نکته را با انتخاب فراپارامترهای کوچک کنترل می‌کنیم.

۴- آزمایشات و نتایج

در مقدمه توضیح داده‌شد که در این مقاله، هدف معرفی مدلی زبانی است که ترتیب کلمات را به‌طورکلی نادیده نمی‌گیرد و درعین‌حال وابسته به ترتیب دقیق کلمات نیست. زیرا در بسیاری از کاربردها بااینکه، ترتیب به‌طورکلی بی‌اهمیت نیست، ولی نقش مستقیم و حیاتی در نتیجه بازی نمی‌کند و به‌این‌ترتیب نیازی نیست تا با در نظر گرفتن ترتیب دقیق کلمات با مشکل تنکی دست‌وپنجه نرم کنیم. در این مقاله، هدف ما این است که یک مدل اولیه را با ویژگیهای مذکور معرفی نموده و آن را از نظر قابلیتش در مدل کردن زبان، با روش‌های n-gram مقایسه نماییم. بنابراین، بررسی و مقایسه مدل پیشنهادی را در کاربردهای ذکر شده به مراحل بعدی کار وامی‌گذاریم.

۴-۱- معیار ارزیابی

در این مقاله، برای ارزیابی مدل از معیار perplexity [۱۴، ۲۴] استفاده شده‌است. این معیار نشان می‌دهد که مدل تخمین زده‌شده تا چه حد، در پیش‌بینی یک نمونه از مدل هدف، خوب است. هرچه مقدار perplexity بدست آمده کمتر باشد، مدل مربوطه مدل بهتری از داده‌ها است. این معیار به‌صورت معکوس میانگین هندسی احتمال کلمات مجموعه براساس مدل، به‌صورتی که در رابطه (۹) آمده‌است محاسبه می‌گردد. در رابطه مذکور w_i یک کلمه در مجموعه داده است، $\mathcal{M}$ نماینده مدل و N نماینده تعداد اسناد مجموعه است. N_m تعداد کلمات هر سند را نشان می‌دهد.

$$ppl = \left(\prod_{i=1}^{N} p(w_i \mid \mathcal{M}) \right)^{-\frac{1}{N}}, N = \sum_{m=1}^{M} N_m \quad (۹)$$

اگر ابتدا لگاریتم رابطه ۸ را گرفته و بعد معکوس آن را انجام دهیم، خواهیم داشت:

$$ppl = a^{-\frac{\left(\sum_{n=1}^{N} log_a\, p(w_{n_{test}} \mid \mathcal{M})\right)}{N}} \quad (۱۰)$$

در رابطه ۹، a می‌تواند مقادیر مختلفی را بپذیرد ولی از آنجا که مدل‌های n-gram و همچنین مقدار perplexity آن‌ها در این مقاله با استفاده از نرم‌افزار SRILM بدست آمده و در این نرم‌افزار مقدار a برابر ۱۰ فرض شده‌است، در این مقاله نیز مقدار آن را ۱۰ در نظر می‌گیریم.

۴-۲- مجموعه‌های داده

در آزمایشات این مقاله، از مجموعه داده 20 newsgroups[۲] استفاده شده‌است. دو نسخه مختلف از این مجموعه داده تشکیل شده‌است. برای ایجاد این دو نسخه، ابتدا، ۲۰۰۰ سند برای یادگیری و ۲۰۰۰ سند برای آزمون، به‌صورت تصادفی، از مجموعه داده‌ها انتخاب شدند. سپس، در نسخه اول که از آن با نام All-Removed یاد خواهیم کرد تمامی ایست‌واژه‌ها، اعداد و علائم حذف شدند. در نسخه دوم که با Substituted نشان داده خواهدشد، ایست‌واژه‌ها، اعداد و علائم هرکدام با یک علامت خاص جایگزین شدند. تفاوت این دو نوع مجموعه داده، این است که در اولی دنباله‌های تکراری بسیار کم هستند و بنابراین داده بسیار تنکی محسوب خواهدشد، در حالی که در دومی، دنباله‌های تکراری زیادی وجود دارند. در بخشهای بعد، نشان خواهیم داد که در هر دو نوع مجموعه، روش پیشنهادی بهتر از مدل‌های n-gram عمل می‌کند. بعد از مقایسه براساس دو مجموعه داده فوق که نشان می‌دهد که مشکل تنکی در مدل پیشنهادی به بزرگی مدل‌های n-gram نیست، مدل‌های مذکور را بر روی یک مجموعه داده متفاوت نیز مقایسه خواهیم کرد. این مجموعه داده Brown[۳] نام دارد و یکی از مجموعه داده‌هایی است که مکرراً در مقالات مربوط به مدل‌های زبانی مورد استفاده قرار گرفته است. مجموعه Brown شامل ۵۰۰ متن ۲۰۰۰ کلمه‌ای از ادبیات انگلیسی است. نتایج حاصل در این بخش نیز نشان می‌دهند که مدل پیشنهادی مدل بهتری از مجموعه داده‌ها است.

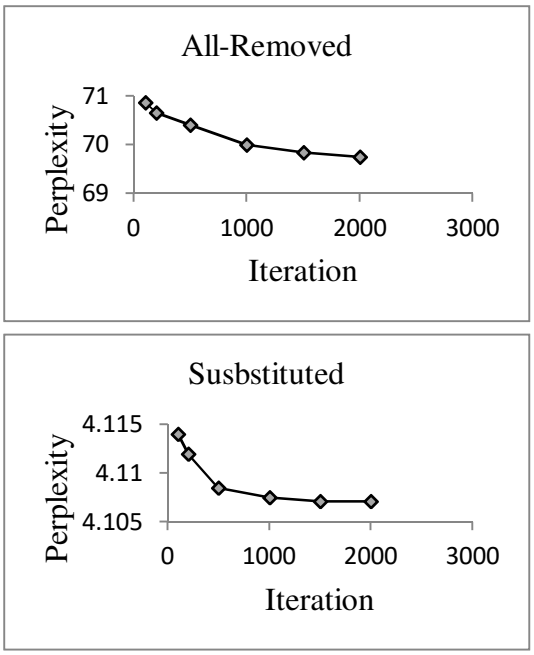
۴-۳- بررسی نتایج حاصل

در مقالات متعددی که به مقایسه مدل‌های n-gram موجود با روش‌های هموارسازی مختلف پرداخته‌اند [۱۴، ۲۴] ذکر شده‌است که روش Knesser-Ney بهبودیافته [۱۴] به‌صورت معناداری بهتر از سایر روش‌های موجود عمل می‌کند و معمولاً روش‌های دیگر با این روش مقایسه می‌شوند. بنابراین در این مقاله نیز روش پیشنهادی با روش

Knesser-Ney بهبودیافته مقایسه شده‌است که از آن با عنوان MKN یاد خواهیم نمود.

شکل ۳ همگرا شدن مدل را بر روی دو زیر مجموعه کوچک شامل ۲۰۰ سند از هر یک از مجموعه داده‌های مذکور را نشان می‌دهد. این آزمایشات با اجرای یک زنجیره مارکوف (نمونه‌برداری گیبس) با ۲۰۰۰ تکرار و پنجره‌ای به طول ۱۰ انجام شده‌اند. همان‌طور که در شکل‌ها قابل مشاهده است کاهش perplexity در ابتدا سریع است و سپس شروع به کندشدن می‌کند. باتوجه به شکل‌ها ۱۰۰۰ تکرار برای حصول نتیجه مناسب می‌تواند کافی باشد.

جدول ۱ نتایج حاصل از اعمال روش MKN برای طول‌های ۱ تا ۵ را روی مجموعه داده‌های مذکور نشان می‌دهد. همچنین این جدول حاوی نتایج حاصل از اعمال روش پیشنهادی برای پنجره‌هایی با طولهای ۲، ۵، ۱۰ و ۲۰ را نشان می‌دهد. در نظر گرفتن پنجره‌ای به طول ۱ به این معنی است که فقط کلمه قبلی می‌تواند به عنوان کلمه سازنده کلمه جاری در نظر گرفته شود. پنجره‌ای به طول ۵ یعنی یکی از ۵ کلمه‌ای که قبل از کلمه جاری قرار گرفته‌اند می‌تواند سازنده آن باشد. سایر طولها را نیز می‌توان به شکل مشابه تفسیر کرد. برای هر کدام از این آزمایشات، ۵ زنجیره مارکوف در ۱۰۰۰ تکرار اجرا شده‌اند و بعد از کنار گذاشتن ۵۰۰ تکرار اول، در هر ۱۰۰ تکرار، ۱ نمونه برداشته شده‌است.



شکل ۳- همگرا شدن مدل پیشنهادی برروی مجموعه داده‌های مورد استفاده

چنان‌که ذکرشد، توقع داریم مدل پیشنهادی عملکرد بهتری داشته باشد زیرا قادر است اثر تنکی را تخفیف دهد. به عنوان مثال، تعداد کلمات یکه در مجموعه داده allRemoved، ۸۳۲۱ است. تعداد عناصر غیرصفر ماتریس φ برای پنجره‌ای به طول ۲۰ برابر ۵۴۸۱۲۶ است که کمتر از ۱ درصد حالت‌های ممکن است. این در حالی است که تعداد 2-gramهای موجود در مجموعه مذکور ۱۳۸۹۶۴ است که کسر خیلی (تقریباً ۴ برابر) کوچک‌تری از حالت‌های ممکن است. پس ماتریس مربوطه در روش پیشنهادی، همچنان تنک است ولی اثر تنکی کاهش یافته است. البته باید توجه داشت که کاهش تنکی صرفاً به معنی افزایش تعداد عناصر غیرصفر نیست، بلکه به‌طورکلی فراوانی رخدادها افزایش می‌یابد و باعث می‌شود مدارک محکم‌تری برای ارتباط کلمات داشته باشیم. از آنجا که ماتریس مربوطه همچنان تنک است، از نظر فضای مصرفی هم، دچار مشکل نخواهیم شد. در جدول ۱، می‌توان مشاهده نمود که روش پیشنهادی با هر چهار اندازه پنجره، بهتر از مدل n-gram عمل نموده است یعنی مقدار perplexity برای روش پیشنهادی کمتر است که نشان می‌دهد روش پیشنهادی هر دو نوع داده را بهتر مدل کرده‌است.

جدول ۱: مقادیر perplexity برای nها و طول پنجره‌های مختلف روی دو مجموعه داده 20 newsgroups

Models	All-removed	Substituted
2-gram	۳۰۵۸.۸۵	۴۰.۹۹
3-gram	۲۸۲۷.۹۱	۴۱.۳۵
4-gram	۲۸۱۹.۳۴	۴۲.۶۹
5-gram	۲۸۲۴.۹۲	۴۴.۱۰
Proposed-model-L2	۳۳۶.۳۸	۱۱.۲۷
Proposed-model-L5	۱۵۵.۱۰	۱۱.۰۰
Proposed-model-L10	۱۱۶.۴۴	۱۱.۰۷
Proposed-model-L20	۱۰۲.۵۳	۱۱.۰۴

جدول ۲: مقادیر perplexity برای nها و طول پنجره‌های مختلف روی مجموعه داده Brown

Models	Brown
2-gram	۴۱۸.۱۸
3-gram	۴۱۹.۹۹
4-gram	۴۲۰.۸۸
5-gram	۴۲۰.۹۰
Proposed-model-L2	۳۷۹.۵۴
Proposed-model-L5	۳۲۳.۸۴
Proposed-model-L10	۳۲۳.۲۷
Proposed-model-L20	۳۱۸.۶۹

روش‌های مذکور همچنین روی مجموعه داده Brown مقایسه شده‌اند که نتایج مربوطه در جدول ۲ گزارش شده‌اند. در این جدول نیز مقدار perplexity برای روش پیشنهادی کمتر است که نشان می‌دهد روش پیشنهادی داده‌ها را بهتر مدل کرده‌است.

۵- نتیجه‌گیری

در این مقاله، یک مدل زبانی بر مبنای مدل‌های سلسله‌مراتبی بیز معرفی‌شده‌است. ایده پایه این مدل این است که در هنگام نگارش یک

متن، کلمه جاری بر مبنای یکی از کلماتی که پیش از آن نوشته شده‌است، شکل می‌گیرد. به عبارت دیگر، یکی از کلمات پیشین است که محرک انتخاب کلمه جاری در ذهن نگارنده می‌گردد. در این مدل، چون کلمات قبل از کلمه جاری مورد توجه قرار می‌گیرند ترتیب تا حدی دارای اهمیت است، اما ترتیب دقیق کلمات دخالت مستقیم در مدل ندارد. به‌همین‌دلیل مشکل تنکی برای این مدل خفیف‌تر از مدل‌های n-gram است. مدل پیشنهادی بر روی دو مجموعه داده متفاوت که هر دو از یک زیر مجموعه ۴۰۰۰ سندی از مجموعه 20newsgroups گرفته شده‌اند با مدل Knesser-Ney که معمولاً در بین مدل‌های n-gram عملکرد بهتری از نظر perplexity دارد مقایسه شده‌اند. دو مجموعه داده فقط از نظر پیش‌پردازش‌های انجام‌شده بر روی آنها متفاوتند. در یکی از آنها تمام ایست‌واژه‌ها، ارقام و علائم حذف شده‌اند و در دیگری هر دسته با علائم یکسان جایگزین شده‌است. روشن است که در حالت اول، دنباله‌های تکراری بسیار تنک هستند ولی در دومی دنباله‌های تکراری بیشتری خواهیم داشت. توانایی مدل پیشنهادی در مدل کردن داده‌های هردو مجموعه با استفاده از معیار perplexity سنجیده شده‌است. مدل پیشنهادی با طول پنجره‌های متفاوت با مدل زبانی Knesser-Ney بهبودیافته [۱۴] با طول‌های مختلف مقایسه شده‌است و در تمام موارد عملکرد مدل پیشنهادی به طور چشمگیری بهتر است. تفاوت مدل پیشنهادی با مدل n-gram در مجموعه اول چشمگیرتر به نظر می‌رسد. این نتیجه منطبق با منطق مدل پیشنهادی است یعنی اثر تنکی تخفیف یافته است. در برخی کاربردها مثل بازیابی اطلاعات، در نظر گرفتن ارتباطات محلی کلمات می‌تواند بسیار مفید باشد اما برای مدل کردن این نوع ارتباطات، لزومی به در نظر گرفتن ترتیب دقیق کلمات نیست. مدل پیشنهادی می‌تواند در چنین کاربردهایی کارآمد باشد. در این مقاله نشان داده‌ایم که مدل پیشنهادی اثر تنکی را کاهش داده است و در مدل کردن داده‌ها بهتر از n-gram عمل کرده‌است، در آینده قصد داریم که عملکرد مدل پیشنهادی را در کاربردهای مذکور بررسی نماییم.

مراجع

[1] Croft, B. and J. Lafferty, "Language modeling for information retrieval. Vol. 13. Springer Science & Business Media, 2013.

[2] Wallach, H.M. "Topic modeling: beyond bag-of-words". ACM.2006.

[3] 3. Manning, C.D., et al., *Introduction to Information Retrieval.* Cambridge University Press. 496. 2008.

[4] Seneff, S. "The use of linguistic hierarchies in speech understanding". in ICSLP. 1998.

[5] Galescu, L. and J.F. Allen. "Hierarchical statistical language models: experiments on in-domain adaptation". in INTERSPEECH. 2000.

[6] Blei, D.M., A.Y. Ng, and M.I. Jordan, "Latent Dirichlet allocation". Journal of machine Learning research. 3(Jan): p. 993-1022. 2003.

[7] Noji, H., D. Mochihashi, and Y. Miyao. "Improvements to the Bayesian Topic N-Gram Models". In EMNLP, pp. 1180-1190. 2013.

[8] Graves, A. and N. Jaitly. "Towards End-To-End Speech Recognition with Recurrent Neural Networks". in ICML. 2014.

[9] Evershed, J. and K. Fitch. "Correcting noisy OCR: Context beats confusion". in Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage. ACM. 2014.

[10] Carlson, A. and I. Fette. "Memory-based context-sensitive spelling correction at web scale. in Machine learning and applications, sixth international conference on. ICMLA. IEEE. 2007.

[11] Kneser, R. and H. Ney. "Improved backing-off for m-gram language modeling". in 1995 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE**.** 1995.

[12] Ney, H., U. Essen, and R. Kneser, "On structuring probabilistic dependences in stochastic language modelling". Computer Speech & Language. 8(1): p. 1-38. 1994.

[13] Jelinek, F. "Interpolated estimation of Markov source parameters from sparse data". in Proc. Workshop on Pattern Recognition in Practice, 1980.

[14] Chen, S.F. and J. Goodman, "An empirical study of smoothing techniques for language modeling". Computer Speech & Language. 13(4): p. 359-394. 1999.

[15] De Mulder, W., S. Bethard, and M.-F. Moens, "A survey on the application of recurrent neural networks to statistical language modeling". Computer Speech & Language. 30(1): p. 61-98. 2015.

[16] Deschacht, K., J. De Belder, and M.-F. Moens, "The latent words language model". Computer Speech & Language. 26(5): p. 384-409. 2012.

[17] Deoras, A., et al., "Approximate inference: A sampling based modeling technique to capture complex dependencies in a language model". Speech Communication. 55(1): p. 162-177. 2013.

[18] Sidorov, G., et al., "Syntactic n-grams as machine learning features for natural language processing". Expert Systems with Applications. 41(3): p. 853-860. 2014.

[19] Deoras, A., et al. "Variational approximation of long-span language models for LVCSR". in 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2011.

[20] Minka, T. and J. Lafferty. "Expectation-propagation for the generative aspect mode"l. Morgan Kaufmann Publishers Inc. 2002.

[21] Griffiths, T.L. and M. Steyvers, "Finding scientific topics". Proceedings of the National academy of Sciences, 101(suppl 1): p. 5228-5235. 2004.

[۲۲] رمضان هاونگی، «موقعیت‌یابی ربات براساس فیلتر ذره‌ای بهبود یافته با فیلتر کالمن گروهی هوشمند و گام MCMC». مجله مهندسی برق دانشگاه تبریز، دوره ۴۶، شماره ۴، صفحه ۳۴۵-۳۵۶. ۱۳۹۵.

[۲۳] سیامک عبداله‌زاده و دیگران، «استفاده از خوشه‌بندی و مدل مارکوف جهت پیش‌بینی درخواست آتی کاربر در وب». مجله مهندسی برق دانشگاه تبریز، دوره ۴۵، شماره ۳، صفحه ۸۹-۹۶. ۱۳۹۴.

[24] Chelba, C., et al., "One billion word benchmark for measuring progress in statistical language modeling". arXiv preprint arXiv:1312.3005, 2013.

زیرنویس‌ها

[1] Vocabulary

[2] http://qwone.com/~jason/20Newsgroups/20news-18828.tar.gz

[3] https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/brown.zip