تاثیر طبقه بندی واژه ای برروی مدل های زبانی

فاطمه ریاحی[1]، هانیه غلامی[2]، ملیحه جعفری[3]، زهرا حبیب فتح آبادی[4]

1 دانشجوی کارشناسی نرم افزار، دانشکده برق و کامپیوتر دانشگاه گناباد، ایران

riyahi1994f@gmail.com

2 دانشجوی کارشناسی نرم افزار، دانشکده برق و کامپیوتر، دانشگاه گناباد، ایران

Just42day92@yahoo.com

3 دانشجوی کارشناسی نرم افزار، دانشکده برق و کامپیوتر، دانشگاه گناباد، ایران

jafari_malihe11@yahoo.com

4 دانشجوی کارشناسی نرم افزار، دانشکده برق و کامپیوتر، دانشگاه گناباد، ایران

m.fathabadi@yahoo.com

چکیده

بازشناسی متون، در سال‌های اخیر بسیار مورد توجه قرار گرفته است .در دهه‌های اخیر با توجه به کاربردهای گسترده مدل زبان، تحقیقات زیادی برای مدل‌سازی زبان‌های پرکاربرد جهانی و به خصوص زبان انگلیسی انجام شده است[1]. این مقاله تاثیر طبقه بندی واژه ای بر روی مدل های زبانی شبکه ای عصبی RNNLM را مورد بررسی قرار می دهد که اخیراً بسیاری از تکنیک های مدل سازی زبانی را معرفی کرده است. این به راحتی برای بهبود تشخیص گفتار موجود و سیستمهای ترجمه استفاده می شود . ما نگاه دقیق تری به این طبقه بندی انجام داده و دریافته ایم که طبقه بندی های پیشرفته می-توانند با ترجمه مناسب عملکرد را نیز بهبود ببخشند. در این مقاله ما در مورد انتخاب پارامتر مطلوب و قابلیت‌های حالت‌های مختلف بحث می‌کنیم به خصوص استفاده از الگوریتم‌های قهوه‌ای را مورد بررسی قرار دادیم که روش کلاسیکی برای طبقه بندی است در آزمایشات استاندارد ما متوجه شدیم که 5 تا 7 درصد از پیچیدگی با استفاده از الگوریتم قهوه ای مسیر است.

واژه های کلیدی: RNNLM، الگوریتم های قهو های، PPL

مقدمه

در کار قبلی ما بسیاری از تکنیک های مدل سازی زبان پیشرفته شناخته شده را ، مقایسه کرده ایم، و متوجه شدیم که شبکه های عصبی مبتنی بر مدل های زبانی (NNLM) بهترین تنظیم های استاندارد را انجام می‌دهند[2].

مدل هایی از این نوع حدود ده سال پیش توسط Bengio معرفی شدند[3]. نقاط ضعف اصلی آنها پیچیدگی محاسباتی بزرگ، و اجرای غیر بدیهی بود. برای کمک به غلبه براین موانع اساسی ،ما تصمیم به انتشار ابزارمون برای آموزش شبکه‌های مبتنی بر مدل‌های زبانی RNNLM داریم.

هدف اصلی از ابزارهای RNNLM به شرح زیر است:

- ارتقاء تحقیقات از تکنیک های مدل سازی زبانی پیشرفته
- استفاده آسان
- کد قابل حمل ساده بدون هیچ وابستگی به کتابخانه‌های خارجی
- بهره وری محاسباتی

یک ویژگی مهم RNNLM یادگیری تاریخچه کل کلمه با لایه‌های مخفی است. به عبارت دیگر لایه مخفی با کلمه حال حاضر و همچنین فعال‌سازی واژه‌های قبلی انجام می‌گیرد. بسیاری از آزمایشات قبلی نشان می دهد که با استفاده از متن های بلند تر می توان مدل زبانی بهتری به دست آورد.

زبان مدل‌سازی آماری توجه هایی را به عنوان مدل‌های زبان های طبیعی که بخش مهمی از بسیاری از سیستم‌های عملی است به خودش جذب می کند.مدل‌های زبانی نحوه توزیع کلمات را در زبان برای عمل بازشناسی گفتار و دیگر فنون زبانی محاسبه می‌کنند.

مدل زبانی یکی از مهمترین اجزای سیستم های تشخیص صحبت ترجمه ماشینی اطلاعاتی و غیره هستند. هدف مدل زبانی مشخص کردن احتمال غیر صفر برای هرگونه توالی واژه در زبان است با نظر گرفتن توالی های کلمات $w_1,w_2,\ldots$ ، احتمال مدل زبانی به شکل زیر است.

$$P(w)= p(w_1^m)= \prod_{i=1}^{i=m} P(w_i \mid w_1^{i-1}) \qquad (1)$$

پیش بینی واژگان فارسی ، یکی از تحقیقاتی است که با استفاده از مدل‌سازی آماری زبان فارسی انجام شده است. در این تحقیق که با وارد شدن حروف اولیه یک واژه ، سیستم واژه هایی را که با آن حرف آغاز می‌شود در پنجره ای فهرست می‌کند و به کاربر پیشنهاد می‌دهد که با وارد شدن حروف بیشتر، پیشنهاد ها محدودتر می‌شود تا واژه مورد نظر کاربر از میان آنها یافت شود.

بررسی یک مدل آماری زبان براساس دسته‌های منطقی دستوری زبان فارسی برای استفاده از بازشناسی گفتار پیوسته ، نوعی مدل‌سازی جدید زبان فارسی است. این مدل زبانی براساس دسته‌های منطقی N-GRAM با طول متغییر کار می‌کند و در آن به جای یافتن الگوهای آماری مربوط به دنباله های کلمات، روابط بین دسته‌های منطقی از کلمات مورد بررسی قرار می‌گیرند[1].

مدل های زبانی N-GRAM نسبت به سایر مدل ها به مراتب درLvcsr مورد استفاده هستند که کلمه پیش بینی شده را بر پایه واژه های N-1 فرض می کنند.

(Large Vocabulary Continuous Speech Recognition)

$$p(w_i|w_1^{i-1}) = p(w_i|w_{i-n+1}^{i-1})$$

کلمات در زبان فارسی می‌توان ترکیبی از زیر کلمات دانست. هر بخش از دنباله حروف که از بخش‌های قبل و بعد بتواند جدا در نظر گرفته شود، یک زیر کلمه است. زیر کلمه از یک حرف یا ترکیبی از حروف به هم پیوسته تشکیل شده است. به عنوان مثال کلمه " مدرسه" دارای سه زیر کلمه "مد"، " ر"، " سه " است. با ترکیب زیر کلمات، کلمات معنی دار ممکن به دست می‌آیند. به عنوان مثال جمله " پدر من آمد " از زیر کلمات " پد" ، " ر" ،" من" ، " آ " ، " مد " تشکیل شده است.از ترکیب ترتیبی (از راست به چپ) این زیر کلمات، تنها می‌توان کلمات معنی‌دار " پدر "، " من " ، " آمد" را استخراج کرد و به همین دلیل ترکیب جمله به راحتی قابل شناسایی است. روند نمای قسمت ساخت کلمات با معنی از ترکیب زیر کلمات در شکل 1 نشان داده شده است. همانطور که در این شکل مشاهده می‌شود ابتدا زیر کلمه اول دریافت می‌شود سپس اگر این زیر کلمه به تنهایی معنی دار باشد، ذخیره و زیرکلمه بعدی از ورودی دریافت می‌شود. همچنین اگر زیرکلمه دریافتی اول بامعنی نباشد، زیر کلمه بعدی از ورودی دریافت می‌شود اگر زیر کلمه دریافت شده نقطه باشد به معنای رسیدن به انتهای جمله و پایان زیرکلمات جمله است. اگر نویسه فاصله باشد به معنای پایان زیر کلمات مربوطه به یک کلمه است. زیر کلمه دریافتی دوم در صورتی که نقطه یا نویسه نباشد، در کنار زیر کلمه اول قرار می‌گیرد و معنی دار بودن ترکیب ساخته شده بررسی می‌شود. در این روند نما M به عنوان یک اندیس برای انتخاب زیر کلمات در نظر گرفته شده است[1].

مدل زبانی N-GRAM ساده و موثر است ولی از لحاظ طبقه بندی دامنه ای گرامری و معنایی بسیار پیچیده است. برای بهبود عملکرد مدل زبانی بسیاری از تکنیک های مدل سازی مناسب معرفی شده اند برخی از آنها از مدل زبانی N-GRAM هم بهتر عمل کرده اند. مانند شبکه های Feed - Forward و ماکزیمم انتروپی و ....

مهم تر از همه شبکه‌های عصبی از نقطه نظر پژوهش بسیار جالب اند همچنین آنها اجازه دارند به پردازش موثر از توالی ها و الگوها با طول دلخواه بپردازند. این مدل‌ها می توانند یاد بگیرند ذخیره اطلاعات گذشته را در لایه‌های پنهانی. با استفاده از الگوریتم قهوه ای طبقات واژه ها دقیق تر بوده و می‌تواند به صورت خودکار انجام گیرد و در این مقاله ما تاثیر طبقه بندی را روی شبکه های عصبی مدل های زبانی قرار داده و از طبقه بندی جدیدی استفاده کرده ایم. آزمایش ها نشان داد که روش های طبقه بندی دقیق تر می تواند عملکرد RNNLM را افزایش دهد.

رابطه زیر مدل N-GRAM

در این مدل، از آمار کلاسیک و احتمال بهره گرفته شده است.به یک توالی n تایی از رأس ها، N-GRAM می گوییم (توالی های 3gram، 2gramو ... داریم). در این مدل، یک مجموعه داده های آماری بسیار بزرگ نیاز داریم که هر کدام مجموعه ای از این نشانه ها به همراه روابط بین آن ها است.روابطی در این مدل تعریف می شود که می توان با استفاده از آن، درستی یک توالی خواص از این نشانه ها را بررسی کرد،می خواهیم درستی عبارت $a_1 \rightarrow a_2 \rightarrow a_3 \rightarrow a_4$ را بررسی کنیم. در این مدل، احتمال درستی به صورت یک عدد تعیین می شود که هرچه داده های آماری ما بیشتر باشد، نتیجه مطلوب تر است. عبارت بالا را به اجزای زیر تقسیم می کنیم و تعداد تکرار هر کدام را در داده های آماری پیدا می کنیم.

$a_1 \rightarrow a_2$

$a_2 \rightarrow a_3$

$a_3 \rightarrow a_4$

$a_1 \rightarrow a_2 \rightarrow a_3$

$a_2 \rightarrow a_3 \rightarrow a_4$

$a_1 \rightarrow a_2 \rightarrow a_3 \rightarrow a_4$

حال احتمال های3-gram، 2-gramو ... به صورت زیر تعریف می شوند:

$P_{2gram} = P(a_2|a_1)\ P(a_3|a_2)$ ... : 2Gram
$P_{3gram} = P(a_3|a_1a_2)\ P(a_4|a_2a_3)$ ... : 3Gram
$P_{4gram} = P(a_4|a_1a_2a_3)$: 4Gram

عبارت $P(a_3|a_1a_2)$یعنی احتمال درستی آمدن a_3 پس از توالی a_1a_2 که مقدار آن برابر عبارت زیر است:

$$P(a_n|a_1 a_2 \dots a_{n-1}) = \frac{C(a_1 a_2 \dots a_n)}{C(a_1 a_2 \dots a_{n-1})}$$

در حالت کلی، احتمال درستی عبارت بالا در n-gram به صورت است:

$$\begin{aligned} P(\mathbf{W}) &= P(a_1, a_2, \dots, a_n) \\ &= P(a_1)P(a_2|a_1)P(a_3|a_1, a_2)\cdots P(a_n|a_1, a_2, \dots, a_{n-1}) \\ &= \prod_{i=1}^{n} P(a_i|a_1, a_2, \dots, a_{i-1}) \end{aligned}$$

(1)

تنها یک مشکل باقی می ماند؛که اگر تنها یکی از این احتمال ها، صفر شود، احتمال کل صفر خواهد شد. در حالی که می دانیم، داده های آماری ما محدودیت دارند و ممکن است، بالاخره، یکی از توالی ها، مخصوصاً اگر تعداد آن زیاد باشد، در داده ها وجود نداشته باشد. راه حل این است که به هر کدام، یک مقدار ثابت (مثلاً 1) اضافه کنیم. روش های زیادی برای رفع این مشکل وجود دارند که در همه آن ها، تابع احتمال به گونه ای تعریف می شود که مقدار صفر برنگرداند.

مدل زبانی عصبی

مدل زبانی عصبی توانسته مشکلات سایر مدل ها را رفع کند.ساختار این روش با طول اختیاری را در شکل 1 میبینیم که شامل یک لایه ی خروجی،یک لایه ی مخفی و یک لایه ی ورودی است.

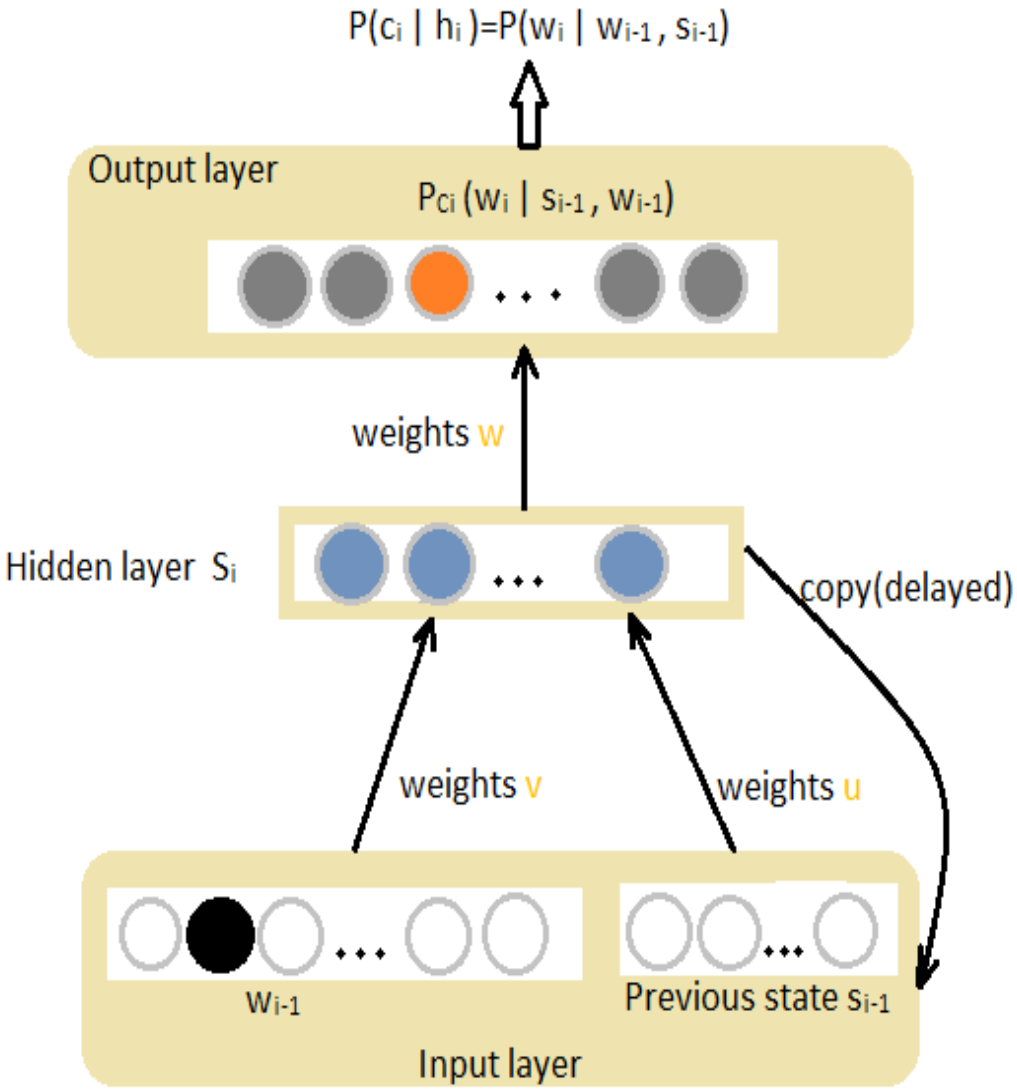

Figure 1 . RNNL Model

شکل 1 مدل زبانی عصبی

لایه ی ورودی شامل کلمه ی آخر و همچنین حالت قبلی میباشد.

عصب هایی که در لایه ی مخفی هستند از یک عملکرد حلقوی پیروی میکنند.لایه ی خروجی نیز احتمال توزیع هریک از کلمات را درحالت قبلی نشان می دهد.
لایه ی ورودی، مخفی و بازده می توانند به شکل زیر باشند:

$$X(t)=w(t)+s(t-1) \quad (2)$$

$$S_j(t)=f(\sum xi(t)uji) \quad (3)$$

$$Yk(t) = g(\sum_j^i \mathrm{Sj(t)Ukj}) \quad (4)$$

$$f(z) = \frac{1}{1+e^{-z}} \quad (5)$$

$$g(Zm) = \frac{e^{zm}}{\sum_k e^{z_K}} \quad (6)$$

برای محاسبه ی خطا نیز از فرمول زیر استفاده میکنیم:

desired یک محور از 1 تا N است که تعداد کلمات در هر متن را نشان میدهد.

$$Error(t)=desired-y(t) \quad (7)$$

CRNNLM

قابل مشاهده است که محاسبه RNNLM بسیار وقت گیر است در این بخش ما ابتدا پیچیدگی محاسبات را بررسی کرده و سپس روش CRNNLM را معرفی می کنیم.

در این تساوی H، اندازه لایه مخفی V ، اندازه کلمه وT میزان مراحل است به طور معمول H بزرگتر V بوده بنابراین پیچیدگی محاسبات بین لایه مخفی و لایه بازده است. برای محاسبات این پیچیدگی بسیاری از روش ها معرفی شده اند اول اینکه تمام واژه های پر تکرار به صورت TOKEN قرار گیرد که سرعت را 2 تا 3 برابر افزایش می دهد. این ایده بعداً گسترش یافت از یک مدل Backoff برای کلمات کم یاب استفاده شود حتی رویکردهای بسیار بهتری نیز طراحی شده اند که همه آنها مقصد بودند که کلمات باید در طبقات گوناگون قرار گیرند.

براساس این ایده مبتنی CRNNLM معرفی شد. مهندسی این روش در شکل 3 نشان داده شده است. CRNNLM که بر پایه RNNLM مبتنی است.

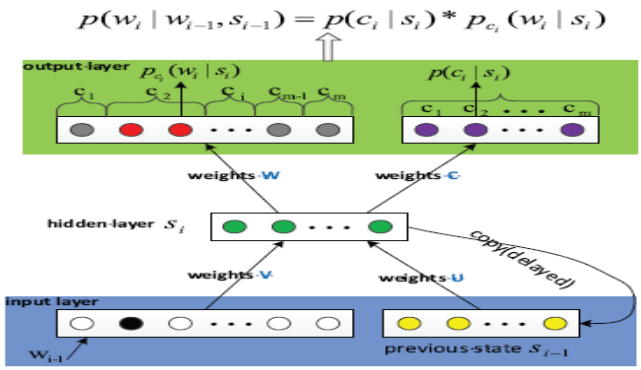

Figure 2. Class based Recurrent Neural Network Language Model

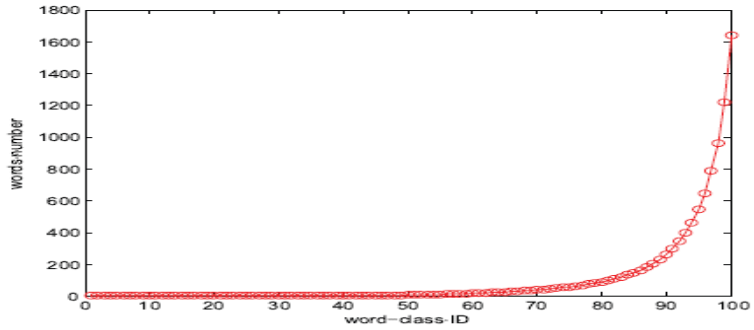

Figure 3. the distribution of word classes given by the frequency-based method. The x-axis is the index of the class and the y-axis is the word number of words, which are contained in the class.

شکل 2 CRNNLM بر پایه RNNLM

استفاده از لایه بازده را به صورت CLASS PART و WORD PART مورد بررسی قرار می دهد. با در نظر گرفتن واژه کنونی Wi که به کلاس Ci متعلق است لایه بازده را Class part احتمال طبقه بندی را نشان داده و به لایه بازده در Word Part احتمال کلمه را نشان می دهد کاربردهای Softmax باید برای تمام بخش هایClass Part مورد استفاده باشد. SOFTPART تنها به کلماتی متعلق است که به مرحله اول پردازش مربوط هستند. نهایتا با در نظر گرفتن کلمات قبلی wi و بازخورد فعالیت مربوط به آن ، احتمال اتفاق کلمات به شکل زیر محاسبه می شوند:

$$p(w_i|w_{i-1}, s_{i-1}) = p(c_i|s_i)p_{c_i}(w_i|s_i) \qquad (9)$$

پیچیدگی محاسبه یک مرحله آموزش CRNNLM نشان داده شده است در این شکل، Cتعداد طبقات است.

$$O = (1 + H) \times H \times \tau + H \times C \qquad (10)$$

به طور معمول تعداد طبقات C از B کمتر است و مهندسی CRNNLM مزایای بهتری نسبت به RNNLM نیز مراحل آزمون و آموزش سریعتری دارد.

روش های طبقه بندی واژه برای CRNNLM

A- روش مبتنی بر فرکانس

روش مبتنی بر فرکانس، ایده اصلی قرار دادن کلمات به صورت متناسب است. به عبارت دیگر کلمات بر اساس احتمالات uni-gram طبقه بندی می شوند. بنابراین اگر بخواهیم کلمات را به 50 طبقه تقسیم کنیم کلمات مرتبط به 2 درصد احتمال در طبقه یک و 2 درصد بعدی در طبقه دوم و ... قرار می گیرند ، بنابراین طبقات اول ممکن است تنها یک کلمه را شامل شوند ولی طبقات آخر ممکن است هزار کلمه مشابهه با شکل دو را داشته باشند. طبقه بندی واژه مبتنی بر فرکانس بسیار ساده ولی غیر دقیق است. استفاده از طبقه بندی های غیر دقیق ممکن است برخی مشکلات را برای مدل زبانی به وجود آورد.

B - الگوریتم قهوه ای:

الگوریتم بالا به پایین در شکل 8 پیشنهاد شده و شامل Tokenهای مختلف بوده و یک درخت باینری را ایجاد می کند. هر مورد داخلی در این درخت می تواند به عنوان یک بخش مورد استفاده قرار گیرد . این طبقه بندی یک حالت سلسله مراتبی دارد.

به صورت مختصر می توان گفت که این الگوریتم ابتدا با تعیین کردن یک Token برای بخش خود عمل کرده و سپس این طبقات را با یکدیگر ادغام می کند تا بتواند کیفیت طبقه بندی را افزایش دهد و کیفیت طبقه بندی نیز با استفاده از مدل زبان طبقه بندی تعریف می شود. دقت داشته باشید که این الگوریتم یک طبقه بندی اولیه را ایجاد می کند زیرا هر Token را به یک طبقه متصل می کند. از دیدگاه معنا گرا، کلمات هم معنی زیادی وجود دارند که از نظر مفهومی و گرامری از یکدیگر جدا بوده و گروه بندی آن ها به ویژگی متن بستگی دارد. طبقه بندی حاصل از الگوریتم قهوه ای مفهوم مدنظر ما را نمی رساند . به علاوه به کارگیری Liang نیز امری معمول است.

در مقایسه با روش مبتنی بر فراوانی ، الگوریتم قهوهای بسیار دقیق تر است . بنابراین ما تصمیم گرفتیم از الگوریتم قهوه ای برای بدست آوردن طبقات استفاده کنیم.

نتایج و کارهای آینده:

در این مقاله ها تصمیم گرفتیم مشخص کنیم که آیا روش های طبقه بندی دقیق تر می تواند عملکرد مدل های زبانی را افزایش دهد یا نه. به ویژه ، تاثیر طبقه بندی واژه را بر روی مدل های زبانی شبکه ای عصبی بررسی کردیم.

روش های مبتنی بر فرآوانی و الگوریتم قهوه ای برای بررسی این عوامل بررسی شدند. نتایج آزمایش نشان داد که کاهش پیچیدگی 5 تا 7 درصدی الگوریتم قهوه ای نسبت به روش دیگر وجود داشت. در این پژوهش، ما تنها روش های اولیه را مورد بررسی قرار دادیم، در کارهای آینده، ما قصد داریم که روش های دقیق تری را کشف کنیم. بسیارخوب است که کلمات متعلق به یک کلاس دیگر را نیز بررسی کنیم که روش طبقه بندی دقیق نامیده می شود. بعلاوه می توانیم از روش هایPart Of Speech نیز سود ببریم.

منابع:

1) بازشناسی متون فارسی با استفاده از مدل زبانی n-gram و پالایش گرامری

2) T. Mikolov, A. Deoras, S. Kombrink, L. Burget, and J.Cernock ´y, "Empirical evaluation and combination of advanced language modeling techniques," in Proceedings of Interspeech , 2011

3) Y. Bengio, R. Ducharme, P. Vincent et al. , "A neural probabilistic language model," Journal of Machine Learning Research , vol. 3, pp. 1137–1155, 2003.

4) ر.روزنفلدو " روش حداکثر آنتروپی به انطباق زبان مدل سازی آماری" گفتار کامپیوتر و زبان ،ج.10 .ص 187، 1996

5) ساعت . شونک، "مدل مستمر زبان فضا" سخنرانی کامپیوتر و زبان ، ج 21 ، نه. 3، 518 -pp.492، 2007

6) بازدید کنندگان . چن و j . گودمن، " یک مطالعه تجربی از تکنیک های مدل سازی برای زبان صاف " در مجموعه مقالات نشست سالانه 34 در انجمن زبان شناسی محاسباتی . انجمن زبان شناسی محاسباتی ، 1996، صص 310-318

7) JT گودمن، "کمی پیشرفت در زبان مدل سازی تمدید نسخه، " تحقیقات مایکروسافت، فن آوری. نماینده MSR -TR- 2001-72، 2001

8) M. هاتر ، "جایزه فشرده سازی دانش بشر ، " 2006

9) R. روزنفلد، "دو دهه از زبان مدل سازی آماری: که در آن انجام ما از اینجا بروم؟ "مجموعه مقالات IEEE، ج 88، صص 1270-1278، 2000

10) A. Stolcke ، "SRILM - یک ابزار مدل سازی زبان توسعه،" در مجموعه مقالات ICSLP، 2002

11) F. JELINEK، "از سه خطیها! مبارزه برای بهبود زبان مدل، مدل ها "در مجموعه مقالات Eurospeech، 1991.

12) http://www.civilica.com/Paper-ICIKT02-ICIKT02_111.html

13) http://Dbase.irandoc.ac.ir

14) http://Www.wikipedia.com