دانشگاه قم

دانشکده فنی و مهندسی

پایان نامه کارشناسی ارشد رشته مهندسی فناوری اطلاعات

عنوان

ارائه روشی جدید برای مدل زبانی تطبیقی

استاد راهنما

دکتر روح الله دیانت

استاد راهنمای دوم

دکتر باقر باباعلی

نگارنده

علی حیدری

اسفند ماه 1393

تشکر و قدردانی

حمد و سپاس خدای را که توفیق کسب دانش و معرفت را به ما عطا فرمود. در اینجا برخود لازم میدانم از تمامی اساتید بزرگوار، به ویژه اساتید دوره کارشناسی ارشد که در طول سالیان گذشته مرا در تحصیل علم و معرفت و فضائل اخلاقی یاری نموده‌اند تقدیر و تشکر نمایم.

از استاد گرامی و بزرگوار جناب آقای دکتر روح الله دیانت که راهنمایی اینجانب را در انجام تحقیق، پژوهش و نگارش این پایان نامه تقبل نموده‌اند تشکر و سپاسگزاری دارم.

از جناب آقای دکتر باقر باباعلی به عنوان راهنمای دوم که با راهنمایی خود مرا مورد لطف قرار داده‌اند کمال تشکر را دارم.

چکیده

در این پایان نامه روشی برای تطبیق مدل زبانی ارائه شده است. این روش، برمبنای ترکیب الگوریتم کاهش بعد Locally Linear Embedding و مدل زبانی n-gram عمل میکند. الگوریتم Locally Linear Embedding در کاهش ابعاد ساختار داده اصلی را حفظ مینماید. لذا انتظار داریم ساختار کلی ماتریس سند-کلمه در این کاهش بعد دچار خدشه زیاد نگردد. الگوریتم ارائه شده، با استفاده از زبان c++ و بهره گیری از توابع موجود در ابزارهای mlpack، armadillo و irstlm پیاده‌سازی گردید. نتایج پیاده‌سازی بهبود حدود 40 درصدی نسبت به n-gram متعارف را نشان میدهد.

واژگان کلیدی: مدل زبانی تطبیقی، الگوریتم Locally Linear Embedding، مدل زبانی

فهرست مطالب

عنوان **صفحه**

فصل اول) کلیات

1–1**)مقدمه** 2

2–1**)مقدمه** 5

2–2**) مروری کلی بر انواع مدل های زبانی** 7

2–3**)مدلهای زبانی** n-gram 8

2–3–1**)** Unigram**،** Bigram **و** Trigram..... 10

2–3–1–1**)** Unigram 10

2–3–1–2**)** Bigram..... 10

2–3–1–3**)** Trigram 11

2–3–2**) هموارسازی**..... 13

2–3–2–1**) هموارسازی جمع با یک** 13

2–3–2–2**) هموارسازی** Witten-Bell..... 14

2–4**) مدلهای زبانی وفقی**.....17

2–4–1**)روشهای مبتنی بر اطلاعات موضوع**..... 19

2–4–1–1**)مدل های مبتنی بر ترکیب**..... 20

2–4–2**)روشهای مبتنی بر دانش معنایی**..... 21

2–4–2–1**) مدل های** LSA **و** PLSA..... 22

2–5**) ارزیابی مدل های زبانی** 25

2–5–1**) نرخ خطای کلمه**..... 25

2–5–2**) آنتروپی متقابل (**CE **) و** 2–5–3**)سرگشتگی(**PP**):**..... 26

2–5–4**)دقت پیش بینی کلمه (**WPA)..... 29

2–5–5) **نرخ نبود واژگان**(Oov)..... 31

3–1) **مقدمه** 34

3–2) **جایگذاری خطی محلی** (LLE) 35

3–3) **ترکیب با** n-gram 38

4–1) **مقدمه** 41

4–2) **تقریب از** n-gram **ایده آل**.....42

4–3) **ترکیب اطلاعات معنایی با اطلاعات** n-gram **متعارف** 44

4–3) **پیشنهاد برای کارهای آتی** 45

منابع انگلیسی.....47

فهرست جدولها

عنوان **صفحه**

جدول (1) **بررسی سرگشتگی تطبیق**.....32

جدول 2) **سرگشتگی روش پیشنهادی در تقریب** n-gram **ایده‌آل** 42

جدول 3) **سرگشتگی روش پیشنهادی در مقابل** n-gram **متعارف (ترکیب اطلاعات معنایی با اطلاعات** n-gram**)** 44

فهرست شکلها و نمودارها

عنوان **صفحه**

شکل 1– **سیستم تشخیص گفتار** 2

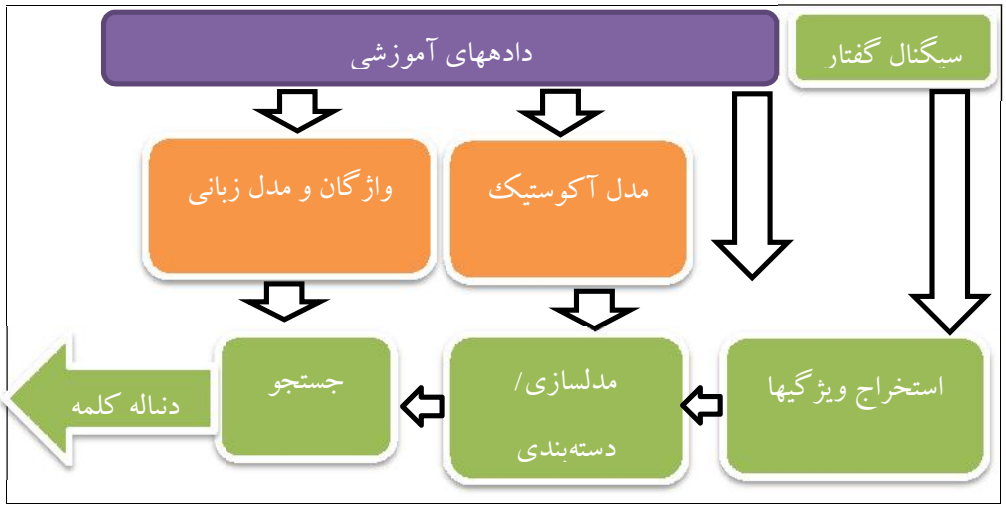
شکل 2– **مدل زبانی سیستم تشخیص شماره** [6]..... 6

شکل 3—**ساختار کلی یک مدل زبانی وفقی** 18
شکل 4– **نمایش سند به صورت برداری** 23
شکل 5– **نمودار سرگشتگی روش پیشنهادی در مقابل سرگشتگی** n-gram **ایده‌آل، تقریب** n-gram **ایده‌آل با استفاده از روش پیشنهادی** 43
شکل 6– **سرگشتگی روش پیشنهادی در مقابل** n-gram **متعارف (ترکیب اطلاعات معنایی)** 45

فصل ۱) کلیات

۱–۱)مقدمه

یکی از زیرشاخه‌های پرداز ش سیگنال، پردازش گفتار است که هدف آن تبدیل توالی گفتار گفته شده توسط گوینده(گان) به نوشتار است. برای رسیدن به این هدف سیستمهای تشخیص گفتار منابع دانشی گوناگونی را بکار میگیرند. با این وجود سیستمهای تشخیص گفتار همچنان از مشکلات متعددی رنج میبرند، که از جمله آنها نویز، عدم دقت گویندگان، وجود شباهت بین برخی واج‌های زبان، همین‌طور هنگام ادای دنباله‌ای از کلمات، به روشنی نمیتوان مشخص نمود یک واج به کدام کلمه از دنباله تعلق دارد و مهمتر از همه اینکه قیود نحوی و معنایی زبان را بسادگی مشابه روشی که انسان‌ها با یکدیگر صحبت میکنند، در سیستم بازشناسی به‌کار برد. همان‌گونه که مشخص است برخی از این مشکلات به زبان و قواعد زبانی مرتبط است. برای حل چنین مسائلی از مدلهای زبانی در سیستم تشخیص گفتار استفاده میگردد. در شکل 1) نحوه ارتباط مدل زبانی با سیستم تشخیص گفتار مشخص شده است.



شکل ۱– سیستم تشخیص گفتار

در مدل زبانی بدون در نظر گرفتن مبنای تشخیص گفتار که ممکن است واج، هجا یا کلمه باشد، برای مشخص نمودن چگونگی ادغام این واحدها از نظر ترتیب، متن و معنا، از محدودیتهایی که توسط مدلهای زبانی اعمال می گردد استفاده مینماییم. [1]

نتیجه آزمایش که توسط Lefevre و همکارانش در سال 2001 صورت گرفت نشان داد که در حدود 8 درصد نرخ خطای تشخیص کلمه در سیستمهای تشخیص گفتار به مدلهای آکوستیک تعلق دارد و 92 درصد را میتوان به مدلهای زبانی نسبت داد. [2] این موضوع توسط آزمایش‌های جداگانه صورت گرفته توسط موسسه ATIS و شخص Bertoldi تایید گردید.[3] بنابراین مدل‌سازی زبان نقش بسیار مهمی در سیستم تشخیص گفتار ایفا مینماید.

دلیل اهمیت مدل زبانی وفقی از آن رو است که زبان طبیعی از جنبه‌های گوناگون خود بسیار متغییر است. از این جنبه ها میتوان نخست همواره در حال تکامل بودن زبان را نام برد. دومین مورد، اینکه آمارهای دنباله کلمه در کاربردهای متفاوت با توجه به موضوع کاربرد متفاوت میباشد. سوم انسان‌ها بگونه ای طبیعی کاربردشان از زبان را برحسب موقعیت تنظیم میکنند. در نهایت سبک گفتار انسان‌ها تحت تاثیر شرایط سیاسی، اقتصادی و ... قرار میگیرد. شاید برخی از این تغییرات جزیی بظاهر بی‌اهمیت باشد ولی همین تغییرات بظاهر کوچک مدل زبانی را در تشخیص دچار اشتباه میکنند. [4] [5].

باتوجه به آنچه در مورد اهمیت مدلهای زبانی وفقی گفته شد، ما در این پایان نامه روشی جدید را برای ساخت یک مدل زبانی وفقی با استفاده از الگوریتم غیر خطی کاهش ابعاد جایگذاری خطی محلی(LLE)[1] ارایه میکنیم. در فصل دوم یک مرور کلی بر مدلهای زبانی و معیارهای ارزیابی مدلهای زبانی خواهیم داشت. در فصل سوم مباحث تئوری روش پیشنهادی، در فصل چهارم نتایج پیاده سازی روش پیشنهادی را توضیح میدهیم.

[1] - locally linear embedding (lle)

فصل ۲) مدلهای زبانی، هموارسازی و ارزیابی مدلهای زبانی

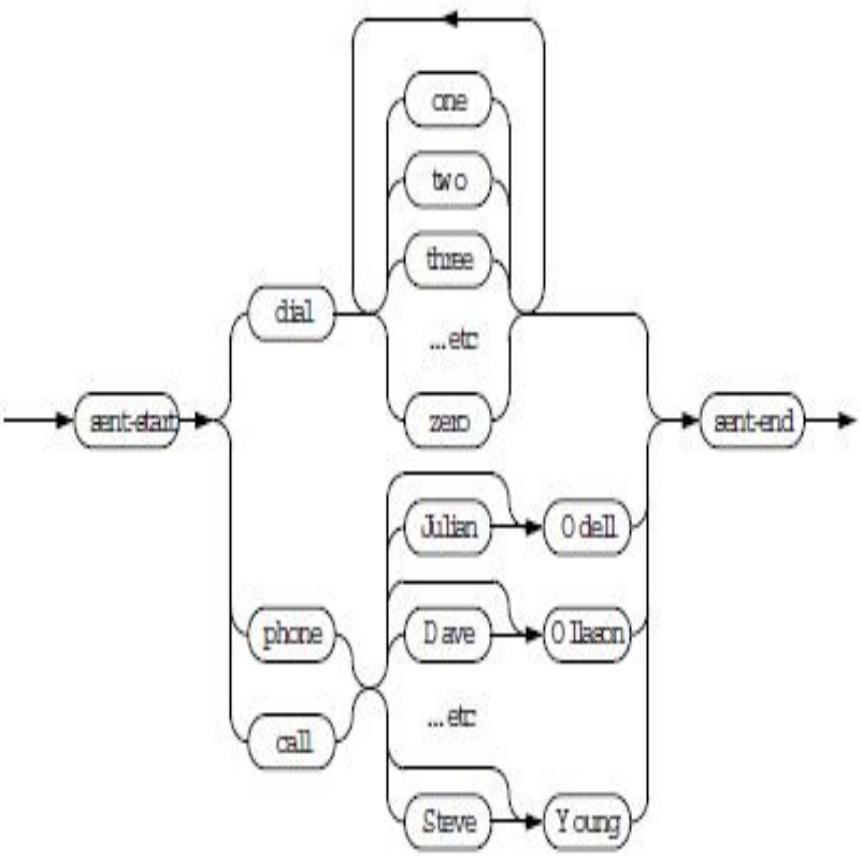
۲–۱)مقدمه

در روشهای بازشناسی گفتار معمولا بجای اینکه صرفا یک کلمه یا یک جمله بعنوان نتیجه بازشناسی ارائه شود چند کلمه یا جمله کاندید ارائه میگردد که ما باید از میان آنها بهترین نتیجه را انتخاب کنیم. بطور مشخص ما از بین تمام جملات کاندید، جمله $W=w_1w_2...w_n$ را میبایست انتخاب کنیم بنحوی که برطبق یک معیار مشخص بهترین جمله باشد. در اینجا مفهوم مدل زبانی به کمک ما می‌آید. به کمک مدل زبانی میتوان تشخیص داد کدام جمله با احتمال بیشتری در زبان مورد استفاده قرار میگیرد. در این صورت بهتر است که به عنوان نتیجه بازشناسی پذیرفته گردد.

مشخصا در یک مدل زبانی عباراتی مثل $P(w_i)$، $P(w_i|w_j)$، $P(w_i|w_j,w_k)$ و ... محاسبه میشود. مثلا $P(w_i|w_j)$ احتمال رخداد عبارت w_iw_j یا w_jw_i را نشان میدهد. برای روشنتر شدن موضوع به مثال زیر توجه نمایید.

مثال:

در یک سیستم بازشناسی گفتار برای شماره گیری تلفن، کاربر ارقام شماره را به ترتیب ادا میکند. سیستم بازشناسی، این ارقام را تشخیص میدهد. در شکل 2) مدل زبانی مربوط به این سیستم تشخیص شماره ارائه شده است.



شکل ۲- مدل زبانی سیستم تشخیص شماره [۶]

این شکل نشان میدهد که در سیستم تشخیص شماره فرض بر این است که کاربر ابتدا یکی از کلمات dial، phone یا call را ادا میکند. در ادامه، کاربر میتواند یکسری ارقام را تلفظ نماید همچنین میتواند نام خانوادگی و یا نام فرد را وارد کند. اگر نتیجه بازشناسی با ساختار گفته شده منطبق نباشد میفهمیم نتیجه بازشناسی اشتباه است. ساختار نشان داده شده در شکل بالا، یک مدل زبانی ساده است که در واقع ساختار یا زبان سیستم را توصیف میکند. در بازشناسی های پیچیده‌تر مثل بازشناسی عبارات زبان فارسی، باز میتوان مدل زبانی درنظرگرفت که ساختار زبان را به نحوی

بیان نماید. درنظر گرفتن چنین ساختاری، کمک مهمی به الگوریتمهای بازشناسی گفتار مینماید. به بیان مشخص در یک مدل زبانی، هدف این است که به ازای هر رشته کلمه $w=w_1w_2...w_n$. (w_i یک کلمه میباشد)، $P(w)$ احتمال رخداد رشته w در زبان را محاسبه نمائیم. معمولا نتیجه یک الگوریتم بازشناسی گفتار به صورت چند رشته میباشد که در نهایت یکی از آن ها باید بعنوان نتیجه نهایی انتخاب شود. با داشتن مدل زبانی میتوان احتمال رخداد هر یک از رشته‌ها را در زبان محاسبه کرد و رشته متناظر با بیشترین احتمال را بعنوان نتیجه نهایی بازشناسی انتخاب نمود.

هدف مشترک همه روشهای ساخت مدل زبانی، محاسبه $P(w)$ میباشد. تفاوت بین روشها در روشی است که برای محاسبه $P(w)$ انتخاب میکنیم.

۲-۲) مروری کلی بر انواع مدل های زبانی

مدلسازی زبانی نه تنها در سیستمهای بازشناسی گفتار، بلکه در زمینه‌های نظیر ترجمه ماشینی، بازشناسی متن، بازیابی اطلاعات و... نیز بکار میروند.مدلهای زبانی آماری از سال 1980 در سیستمهای تشخیص گفتار بکار میرود.

در [7] مدلهای زبانی آماری به پنج دسته تقسیم گردیده‌اند. این پنج دسته عبارتند از : 1) n-gram ها، 2) مدلهای برانگیخته شده توسط قواعد زبانی، 3) مدلهای مبتنی بر درخت تصمیم، 4) مدلهای نمایی، 5) مدلهای وفقی. n-gram که ساده ترین مدل زبانی است. اولین بار توسط شانون در 1948 ارائه گردید [8]. باید توجه داشت که زبان طبیعی دارای ساختارهای پیچیده‌تر از روشی است که مدلهای n-gram ارائه میدهند، از این رو مدلهای زبانی مبتنی بر ساختارهای نحوی و مبتنی بر معنا و مفهوم زبان برای جبران نقایص n-gram ها پیشنهاد گردیده‌اند.

در ارتباط با مدلهای انگیخته شده توسط قواعد زبان میتوان به گرامرهای مستقل از متن[۲] که در [9] پیشنهاد گردیده نام برد. همچنین میتوان مدلهای ساختیافته ارائه شده توسط Chelba و Jelinek اشاره نمود [10] . این مدلها درپی یافتن بهترین درخت تجزیه برای جمله یا بخشی از آن با استفاده از تحلیلگرهای نحوی پایین به بالا[۳] هستند. تحلیلگرهای نحوی جملات را مبتنی بر قواعد نحوی تجزیه میکنند. مزیت این روشها در استفاده از سوابق دورتر برای تخمین آماره کلمه مورد پیش بینی است. مدلهای مستقل از متن به علت پیچیدگی در پیاده‌سازی تقریبا فاقد کاربرد هستند.

مدل های زبانی مبتنی بر درخت تصمیم اولین بار در [11] ارائه شده‌اند. این مدلها فضای تاریخچه را با استفاده از پرسشهای دودویی درباره تاریخچه در گرههای داخلی درخت بخش بندی میکنند، سپس در هر برگ از دادهای آموزشی برای ساخت توزیع احتمال کلمه بعدی استفاده میکند.

مدلهای نمایی منابع دانش مختلف را به صورت یک سری محدودیت خطی (قید) به گونه‌ای که توزیع احتمال مورد نظر در آن قیدها صدق کند فرموله بندی میکنند. از جمله مدلهای نمایی میتوان به برآورد حداقل اطلاعات تفاضلی (MDI)[۴] که توسط Della Pietra و همکارانش در 1992 ارائه گردید اشاره نمود [12] . در ادامه مهمترین مدلهای زبانی n-gram ، مدلهای زبانی وفقی و همینطور برخی مدلهایی که با مدلهای وفقی مرتبط هستند، تشریح خواهند شد.

۲-۳)مدلهای زبانی n-gram

یک مدل زبانی n-gram، در اصل برای محاسبه احتمال P(W) به کار میرود. P(W) نشان دهنده احتمال رخداد رشته W در زبان مورد نظر است. به عنوان مثال، ما میتوانیم داشته باشیم :

[2] - Context free grammar(CFG)
[3] -Bottom UP
[4] - Minimum Discrimination Information(MDI)

P(hi)=0.01، از سوی دیگر ، ما داریم : P(lidgallops chang sha ppp)=o زیرا چنین جمله ای اصولا در زبان انگلیسی وجود ندارد.

مطابق آنچه در احتمال خوانده ایم، P(w) می تواند به صورت زیر تجزیه شود.

(1)

$P(w)= P(w_1....w_2,.....w_n) = P(w_1)\ P(w_2|w_1)\ P(w_3|w_2)\ ...P(w_n\ |w_1w_2.....w_{n-1}) = \pi_{i=1}^{n}\ P(W_i\ |W_1,W_2,.....,W_{i-1})$

که $P(W_i\ |W_1,W_2,.....,W_{i-1})$ ، احتمال آنکه وقوع w_i به شرط دنباله ی $W_1,w_2,.....w_{i-1}$ قبل از آن رخ داده باشد را بیان میکند. به عبارتی دیگر خواهیم داشت:

(2)

$$P(w) = \prod_{i=1}^{v} P(w_1,w_2,...,w_{i-1})$$

یک مدل زبانی که احتمال وقوع کلمه فعلی را به n کلمه بلافاصله قبلیش وابسته می نماید، مدل n-gram است . در چنین مدلی داریم [13] .

(3)

$$P(w) = \prod_{i=1}^{n} P(w_1,w_2,...,w_{i-n+1})$$

مقدار بزرگتر از 3 برای n از نظر عملی مناسب نیستند به دلیل اینکه ترکیباتی بالای 3 نیاز به حافظه و محاسبات زیادی دارند . درنتیجه ما معمولا فقط مدلهای Unigram ، Bigramو Trigram را داریم.

۲-۳-۱) Unigram، Bigram و Trigram

در ادامه مدلهای n-gram، Unigram، Bigram و Trigram که مهمترین و کاربردیترین مدلها برای ساخت مدلهای زبانی هستند، توضیح دادهمیشوند.

۲-۳-۱-۱) Unigram

ساده ترین نوع n-gram است که تنها وابسته به کلمه جاری بوده و به توالی کلمات ماقبل وابستگی ندارد[14] . احتمال رخداد یک کلمه در این مدل بصورت زیر است :

(4)

$$P(w) = \prod_{i-1}^{n} P(w_1)$$

۲-۳-۱-۲) Bigram

اگر رخداد یک کلمه به کلمه ماقبل وابسته باشد ما مدل زبانی Bigram را خواهیم داشت. در حالت $i=1$ برای معنادار ساختن $p(W_i|W_{i-1})$، ما بطور دستی کارکتر(توکن)$<s>$ را بعنوان شروع جمله در نظر میگیریم:

(5)

$$P(w) = P(w_1| < s >) \prod_{i=2}^{n} P(w_i|w_{i-1})$$

همچنین واضح است که در انتهای جملات نیز برای آن که مجموع احتمالات همه ی رشته ها برابر 1 شود ، لازم است توکن انتهای دستی $|s>$ (بعنوان کارکتر انتهایی عبارت) در انتهای جمله قرار داده شود. بعنوان مثال فرض کنید میخواهیم P(mary loves that person) ، را محاسبه کنیم.

P(mary loves that person)=p(mary | <s>)p(loves | mary)
p(that|loves)p(person | that)p(<|s> | person)

Trigram (۳-۱-۳-۲

مدل n-gram دیگر مدل Trigram است که در آن احتمال رخداد یک کلمه وابسته به دو کلمه ی قبلش است. برای مدل زبانی Trigramمحاسبه امتیاز مدل زبانی بصورت رابطه زیر است.

(6)

$$P(w) = P(w_1| < s >)P(w_2| < s >, w_1)\prod_{i=3}^{n} P(w_i|w_{i-2}, w_{i-1})$$

همانطور که می بینید فضای جستجو در مدل Trigram پیچیده تر می باشد و مشخص است که پیاده سازی جستجوی Trigram با واژگان بزرگ پر هزینه است. بنابراین نیاز داریم که گراف جستجوی Trigram را بصورت پویا با استفاده از الگوریتم جستجوی گرافی تولید کنیم .

یکی مشکل عمده در مدل سازی n-gram ، پراکندگی داد ه ی آموزشی میباشد، اگر پیکره متنی آموزشی به اندازه کافی بزرگ نباشد، برخی ازتوالی‌های کلمه ممکن است عملاً به خوبی مشاهده نشوند که منجر به برخی احتمالات بسیار کوچک می شوند. به طور مثال، با مجموعه های چند میلیون کلمه ای از متن انگلیسی، بیش از 50٪ trigram ها، تنها یک بار رخ میدهند و بیش از 80٪ trigram ها کمتر از 5 بار رخ میدهند. هموار سازی برای قوی تر ساختن احتمالات تخمین زده شده، برای داد ه ی دیده نشده، مهم است [15] . در زیر با مثالی ساده‌تر نحوه محاسبه مدل زبانی bigram مبتنی بر کلمه و ضعف این مدل در برآورد صحیح را نشان میدهیم:

JOHN READ MOBY DICK

MARY SHE READ A DEFERENT BOOK

SHE READ A BOOK BY CHER

برای مثال فوق خواهیم داشت.

(7)

P(JOHN READ A BOOK)
= P(JOHN| < S >) P(READ | JOHN) P(A | READ) P(BOOK | A) P(</S > | BOOK)

$$= \frac{c(< S > JOHN)}{\sum_w c(< S > w)} \frac{c(JOHN\ READ)}{\sum_w c(JOHN\ w)} \frac{c(READ\ A)}{\sum_w c(READ\ w)} \frac{c(A\ BOOK)}{\sum_w c(A\ w)}$$

$$\frac{c(BOOK\ w)}{\sum_w c(BOOK\ w)} = \frac{1}{3}\frac{1}{1}\frac{2}{3}\frac{1}{2}\frac{1}{2} \approx 0.06$$

اما مشکلی که مدل فوق از ان رنج می برد در زیر نشان داده شده است.

(8)

P(CHER READ A BOOK)
= P(CHER| < S >)P(READ | CHER)P(A | READ)p(BOOK | A)P(</S > | BOOK)

$$= \frac{c(< S > CHER)}{\sum_w c(< S > w)} \frac{c(CHER\ READ)}{\sum_w c(CHER\ w)} \frac{c(READ\ A)}{\sum_w c(READ\ w)} \frac{c(A\ BOOK)}{\sum_w c(A\ w)}$$

$$\frac{c(BOOK\ w)}{\sum_w c(BOOK\ w)} = \frac{0}{3}\frac{0}{1}\frac{2}{3}\frac{1}{2}\frac{1}{2} \approx 0$$

در مدل های n-gram، به دلیل کمبود حجم داده‌های اموزشی بسیاری از احتمالات مساوی صفر می باشند. این موضوع اساس هموارسازی است.

۲-۳-۲) هموارسازی

هموارسازی ناظر بر تعدیل نمودن احتمالات پایینتر با کاستن از احتمالات بسیار بالا و افزودن به احتمالات صفر است، بگونه‌ای که توزیع احتمالات یکنواخت گردیده و داده‌های پرت حذف گردند.

تاکنون روشهای هموارسازی متعددی ارائه گردیده‌اند، از جمله این روشها میتوان به هموارسازی Additive، درونیابی Jelinek-Mercer ، روش عقبگرد Katz، هموارسازی Witten-Bell[5] و هموارسازی Kneser-Ney و ... اشاره نمود.

در ادامه روش هموارسازی WB که در این پایان نامه برای هموارسازی مدلهای n-gram استفاده گردیده را توضیح میدهیم. برای درک بهتر این روش ابتدا روش هموارسازی جمع با یک[6] را تشریح میکنیم.

۲-۳-۲-۱) هموارسازی جمع با یک

یک روش ساده که زیرمجموعه روش هموارسازی Additive است، افزودن عدد یک به آمارهای n-gram است.

(9)

$$P(W_{i-n+1} \dots W_{i-2}W_{i-1}) = \frac{N(W_{i-n+1} \dots W_{i-2}W_{i-1}) + 1}{N(W_{i-n+1} \dots W_{i-2}W_{i-1}) + V}$$

در رابطه فوق V تعداد کل واژگان موجود در پیکره است.

[5] - Witten-Bell(WB)
[6] - add-one