

ارائه یک روش جدید بازیابی اطلاعات مناسب برای متون حاصل از بازشناسی گفتار

روح الله دیانت^۱، مرتضی علی احمدی^۲، محمد یحیی اخلاقی^۳ و باقر باباعلی^۴

^۱ استادیار گروه مهندسی فناوری اطلاعات، دانشکده فنی و مهندسی، دانشگاه قم، قم، ایران

^۳ گروه علوم کامپیوتر، دانشگاه خاتم النبیین، کابل، افغانستان

^۴، پردیس علوم، دانشکده ریاضی، آمار و علوم کامپیوتر، دانشگاه تهران، تهران، ایران



چکیده

در این مقاله، یک پیش پردازش روی روش‌های بازیابی اطلاعات، ارائه می‌شود که برای بازیابی اطلاعات حاصل از متون بازشناسی شده گفتاری، مناسب است. این پیش پردازش، به شکل ترکیبی از اصلاح و گسترش پرس و جو است. ورودی‌های مسئله، اسناد متنی به دست آمده از بازشناسی گفتار و پرس و جو است و هدف، یافتن اسناد مرتبط با کلمه پرس و جو است. مشکل آن است که متن حاصل از بازشناسی گفتار، همواره دارای درصد خطایی در بازشناسی است که ممکن است منجر به این شود که کلماتی که در واقع مرتبط هستند و به علت وقوع خطای بازشناسی دگرگون شده‌اند، مرتبط تشخیص داده نشوند.

ایده روش ارائه شده، تشخیص خطای بازشناسی در کلمات و در نظر گرفتن کلمات مشابه برای آن دسته از کلماتی است که به عنوان خطا تشخیص داده شده‌اند. برای تشخیص کلمه خطا، پارامتری به عنوان احتمال خطا در کلمه تعریف می‌شود که بزرگ بودن آن بیانگر امکان بیشتر وقوع خطا در کلمه است. همچنین برای تشخیص کلمات مشابه، ابتدا با استفاده از معیار فاصله لونشتاین، کلمات مشابه اولیه را پیدا می‌کنیم؛ سپس احتمال تبدیل این کلمات مشابه به کلمه پرس و جو اصلی، محاسبه می‌شود. کلمات مشابه معنایی، از بین کلماتی که احتمال تبدیل بیش تری دارند، بر اساس یک سطح آستانه انتخاب می‌شوند. اکنون در الگوریتم بازیابی، علاوه بر کلمه اصلی، کلمات مشابه آن نیز در جستجو، مرتبط در نظر گرفته می‌شوند. نتایج پیاده‌سازی‌ها نشان می‌دهد که الگوریتم ارائه شده، معیار F را به میزان حداکثر ۳۰٪ بهبود می‌بخشد.

واژگان کلیدی: بازیابی اطلاعات، بازشناسی گفتار، سند، پرس و جو، فاصله لونشتاین

Introducing a New information Retrieval Method Applicable for Speech Recognized Texts

Rouhollah Dianat¹, MortezaAli Ahmadi², MohammadYahya Akhlaghi³ & Bagher Babaali⁴

¹ Assistant Professor in Department of Information Technology Engineering, Faculty of Technical and Engineering, University of Qom, Qom, Iran

² Master graduate, Faculty of Technical and Engineering, University of Qom, Qom, Iran

³ Lecturer in Department of Computer Science, University of Khatam Al Nabeyin, Kabol, Afghanistan

⁴ Assistant Professor, College of Science, Faculty of Mathematics, Statistics and Computer Science, University of Tehran, Tehran, Iran

Abstract

In this article a pre-processing method is introduced which is applicable in speech recognized texts retrieval task. We have a text corpus, t generated from a speech recognition system and a query as inputs, to search queries in these documents and find relevant documents. A basic problem in a typical speech recognized text

is some error percentage in recognition. This, results erroneously assigning to irrelevant documents. The idea of this proposed method, is to detect error-prone terms and to find similar words for each term. A parameter is defined which calculates the probability for occurring errors in the error-prone words. To recognize similar words for each specific term, based on a criterion called average detection rate (ADR) and levenshtein distance criterion, some candidates are chosen as the initial similar words set. And then, a conversion probability is defined based on the conversion rate (CR) and the noisy channel model (NCM) and the words with higher probability based on a threshold level are selected as the final similar words. In the retrieval process, these words are considered in the search step in addition to the base word. Implementation result shows a significant improvement up to 30% of F-measure in information retrieval method with consideration of this pre-processing.

Keyword: Information retrieval, Speech recognition, Document, Query, Levenshtein Distance

یا مرتبطترین سندها به پرس و جوی داده شده را به ترتیب رتبه^۶، به کاربر برمی گرداند.

در بازیابی صدا، پرس و جو می تواند به دو شکل کلی داده شود [19]. در شکل نخست که به آن پرس و جو با نمونه گفته می شود، از یک نمونه صوتی به عنوان پرس و جو استفاده می شود. در این حالت هدف کاربر این است که پوشه های صوتی را که شامل نمونه صوتی ورودی یا نزدیک به آن هستند، پیدا کند. بسیاری از سامانه های بازیابی موسیقی و بازیابی گفتار آهنگین از این روش استفاده می کنند. در حالت خاصی از این روش، کاربر یک آهنگ یا موسیقی را زمزمه می کند و سامانه بازیابی، موسیقی یا آهنگ های مشابه با آن را از یک پایگاه داده یافته و فهرست آن ها را به کاربر برمی گرداند. به این روش پرس و جو با زمزمه^۷ می گویند [9].

در شکل دوم که به آن پرس و جو با کلمه کلیدی^۸ گفته می شود، پرس و جو، یک نمونه متنی شامل یک یا چند کلمه کلیدی است. در این روش که بیشتر در بازیابی صداهای گفتاری به کار می رود، هدف کاربر این است که پوشه های گفتاری را که شامل آن کلمات کلیدی یا در موضوعی نزدیک به آن ها هستند، پیدا کند. در حالتی که با صداهای غیر گفتاری سرو کار داریم، پرس و جوی متنی می تواند ویژگی ها یا توصیفی از صوت باشد که در قالب کلمات بیان می شود.

برای بازیابی صداهای گفتاری به طور معمول از ترکیبی از فناوری های بازشناسی گفتار^۹ و بازیابی اطلاعات استفاده می شود؛ یعنی سندهای گفتاری ابتدا به متن معادل تبدیل شده و سپس عملیات بازیابی بر روی متن انجام می شود [18]. در شکل زیر شمای کلی یک سامانه بازیابی صدای گفتاری نمایش داده شده است:

۱- مقدمه

اگر چه امروزه استفاده از سامانه های رایانه ای، جستجو و بازیابی اطلاعات^۱ را تسهیل کرده است؛ ولی جستجو و بازیابی، تاکنون به طور عمده در محتوای متنی انجام می پذیرفته است؛ در حالی که نیاز به استخراج اطلاعات از محتوای غیر متنی مانند صدا نیز احساس می شود، هر چند این نیاز تا حدی با تولید فراداده^۲ برای محتوای صوتی رفع می شود، ولی تولید فراداده کاری زمان بر و پرهزینه است. بنابراین به دنبال روش یا روش هایی هستیم که بدون نیاز به فراداده و تنها با استفاده از محتوای خود صدا و مشخصات آکوستیکی آن، جستجو و بازیابی اطلاعات را در آن انجام دهد. به این روش ها، بازیابی مبتنی بر محتوا^۳ گفته می شود که بازیابی صدا یکی از این موارد محسوب می شود.

مسئله بازیابی صدا را می توان به صورت زیر تعریف کرد: یک مجموعه از مستندات صوتی^۴ در اختیار داریم؛ می خواهیم از بین این مجموعه مستندات، سند یا سندهایی را پیدا کنیم که درجه شباهت و یا میزان ارتباط و نزدیکی آن با یک نمونه صدای مورد نظر (و یا توصیفی از آن صدا) بیشتر باشد.

نمونه صدا (یا توصیفی از آن) را که به دنبال سندهای مرتبط با آن در مجموعه هستیم، در اصطلاح پرس و جو^۵ نامیده می شود. با این تعریف، یک سامانه بازیابی صدا به طور کلی به این صورت عمل می کند که کاربر درخواست خود را به صورت پرس و جو به سامانه می دهد. سامانه بازیابی صدا با روش های خاص خود میزان شباهت و یا میزان ارتباط و نزدیکی پرس و جوی داده شده را، با تک تک سندهای صوتی موجود در پایگانی می سنجد و سپس فهرستی از شبیه ترین و

¹ Information retrieval

² Metadata

³ Content base retrieval

⁴ Audio documents

⁵ Query

⁶ Rank

⁷ Query by humming

⁸ Query by keyword

⁹ Speech recognition

مروری بر روش‌ها و کارهای مشابه در این زمینه خواهیم داشت. بخش چهارم به توضیح الگوریتم‌های پیشنهادی اختصاص دارد. نتایج پیاده‌سازی و تحلیل‌های مربوطه در بخش پنجم آورده شده است و در نهایت، بخش ششم به نتیجه‌گیری و ارائه پیشنهادها برای انجام کارهای آینده اختصاص دارد.

۲- پیشینه پژوهش

در مقدمه مقاله حاضر، توضیحاتی در مورد مفهوم بازیابی صداهای گفتاری ارائه و بیان شد که تمرکز مقاله حاضر، روی بازیابی صداهای گفتاری است که به متن تبدیل شده‌اند. مشکل عمده‌ای که در این‌جا وجود دارد، وجود خطای بازشناسی گفتار است که موجب می‌شود سندی که واقعاً مرتبط است، نامرتب تشخیص داده شود. مشابه این مشکل، در بازیابی متون عادی نیز ممکن است در اثر مواردی چون وقوع خطاهای املایی یا تایپی، اتفاق بیافتد.

یک راه‌کار کلی برای حل این مشکل، گسترش پرس‌وجو^۱ است. در این روش، در کنار کلمه اصلی پرس‌وجو، تعدادی کلمه مشابه، در نظر گرفته می‌شود و اسنادی که شامل این کلمات باشند نیز جزو اسناد مرتبط در نظر گرفته می‌شوند. مشابهت در این‌جا می‌تواند از نوع معنایی و مترادف بودن و یا ساختاری باشد. علاوه بر گسترش پرس‌وجو، روش‌های اصلاح پرس‌وجو نیز که به‌طور معمول به‌صورت تصحیح خطاهای تایپی یا املایی انجام می‌شود نیز مورد استفاده قرار گرفته است.

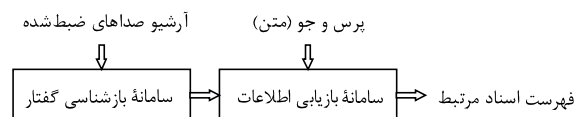
در حوزه بازیابی متون حاصل از بازشناسی گفتار، از بین دو روش گسترش و اصلاح پرس‌وجو، در بیشتر موارد، روش نخست استفاده شده است. یافتن کلمات مشابه نیز در برخی مقالات، بر مبنای استفاده از قواعد نحوی (مانند [4]) و در بعضی دیگر بر مبنای ویژگی‌های آکوستیکی (مانند [14]) انجام گرفته است.

علاوه بر دو روش یادشده، روش‌های دیگری نیز ارائه شده است. به‌عنوان مثال، صرف‌جو در [2]، روشی بر مبنای گسترش اسناد ارائه کرده است. در این روش، ابتدا کلمات مهم و کلیدی هر سند، شناسایی و سپس، به‌ازای هر کلمه، تعدادی کلمه مشابه، استخراج می‌شوند. بدین ترتیب، در واقع به‌ازای هر سند، تعدادی سند مشابه به دست آمده است. به‌علت تعداد بسیار زیاد این اسناد مشابه، در مرحله بعد، تعداد محدودی از مناسب‌ترین آن‌ها انتخاب می‌شوند و

در این مقاله، تمرکز ما روی همین سامانه است، یک مشکل عمده این است که سامانه‌های بازشناسی گفتار همواره با درصدی خطا همراه هستند. در نتیجه ممکن است، سندی مرتبط با پرس‌وجو باشد، اما به‌علت خطای بازشناسی گفتار و در نتیجه وجود کلمات دچار اشکال در متن، به‌عنوان سند مرتبط تشخیص داده نشود.

در این مقاله، روشی جدید برای بازیابی اطلاعات گفتاری بر اساس گسترش پرس‌وجو ارائه شده است. در این‌جا، ایده این بوده که روش بازیابی نسبت به رخداد خطای بازشناسی گفتار تا اندازه‌ای مقاوم باشد. برای پیاده‌سازی، انجام سه گام، پیشنهاد شده است.

گام نخست، تشخیص احتمال خطای کلمات است. گام دوم، یافتن کلمات مشابه کلمه مورد پرس‌وجو است؛ یعنی کلمات جایگزینی را که ممکن است پرس‌وجو بعد از بروز خطا به آن‌ها تبدیل شود، شناسایی می‌کنیم. نکته مهم این است که در این‌جا فقط یافتن کلمات مشابه کافی نیست؛ بلکه از آن‌جا که یک کلمه بعد از بروز خطا با احتمال متفاوت به کلمات دیگر تبدیل می‌شود، محاسبه احتمال تبدیل نیز اهمیت زیادی دارد. گام سوم، بازیابی اسنادی که دچار خطای بازشناسی شده‌اند با استفاده از یک روش بازیابی اطلاعات است. این کار با اضافه کردن پارامترهای به‌دست‌آمده در دو گام قبلی، به یک روش بازیابی صورت می‌گیرد.



(شکل-۱): ساختار کلی یک سامانه بازیابی اطلاعات گفتاری با

استفاده از کلمات کلیدی

(Figure-1): The general structure of a speech information retrieval system using keywords

برای تحقق و پیاده‌سازی اجزای مختلف این ایده، الگوریتم‌هایی پیشنهاد شده است که به‌زودی توضیح داده خواهند شد.

کلیه بخش‌های روش پیشنهادی به‌طور کامل پیاده‌سازی شده‌اند. نتایج حاصل، بیان‌گر کارایی بهتر این الگوریتم نسبت به الگوریتم‌های معمول بازیابی اطلاعات است. بخش‌های مختلف این مقاله به شرح زیر ساماندهی شده است: در بخش دوم پیشینه پژوهش و در بخش سوم

^۱Query expansion

که در آن $w \in D$ کلمه صحیح مورد نظر است و $s \in \Sigma^*$ کلمه اشتباه است.

به‌عنوان مثال، برای حروف الفبای انگلیسی، $\Sigma = \{a, b, c, \dots, y, z, A, B, \dots, Z, \dots\}$ یک زیرمجموعه از Σ^* ، واژگانی از کلمات معتبر انگلیسی خواهد بود. اشتباهات و خطاهای مختلفی وجود دارد که شامل موارد مهم زیر است.

۱- حذف ناخواسته حروف (برای مثال letter به‌جای letter)، اضافه‌شدن تصادفی حروف (برای مثال، mistake به‌جای mistake)

۲- جابه‌جایی حروف (برای مثال، recieved به‌جای received)

۳- جانشینی حروف (برای مثال، fimate به‌جای finite)

برای ساخت ماتریس کانال نوفه‌ای Γ ، ما باید احتمال هر اشتباه (یعنی $\Pr(s|w)$) را به ازای تمامی $w \in D$ و $s \in \Sigma^*$ داشته باشیم. این احتمال ممکن است از طریق فاصله لونشتاین بین w و s محاسبه شود.

هدف مدل کانال نوفه‌ای، یافتن کلمه صحیح از بین کلمات اشتباه ورودی است. برای انتخاب کلمه صحیح، نیاز به یک تابع تصمیم γ است. تابع تصمیم $\sigma: \Sigma^* \rightarrow D$ ، تابعی است که کلمات اشتباه را گرفته و کلمه صحیح را برمی‌گرداند. روش‌های ساخت این تابع تصمیم شامل شیوه بیشترین احتمال (ML^8)، شوهی بیشترین احتمال پسین (MAP^9) و روش کمترین فاصله ویرایشی 10 است.

در یک سیستم احتمالی، ما می‌خواهیم $\hat{w} = \operatorname{argmax}_w P(w|s)$ را بیابیم. به بیان دیگر، در این جا کلمه اشتباه s را در اختیار داریم. می‌خواهیم از بین کلیه کلمات موجود در فرهنگ لغت، کلمه‌ای را انتخاب کنیم که موجب شود این احتمال، بیشینه شود. این کلمه را \hat{w} می‌نامیم. بر اساس قانون بیز 11 ، می‌دانیم رابطه زیر وجود دارد:

$$P(w|s) = P(s|w)p(w) \quad (1)$$

در مدل پیشنهادی، فرض شده است، کلمه‌ای را با احتمال $P(w)$ به‌عنوان پرس‌وجو به سامانه بازبایی،

در نهایت این اسناد، به همراه اسناد اصلی، در فرایند بازبایی، شرکت می‌کنند.

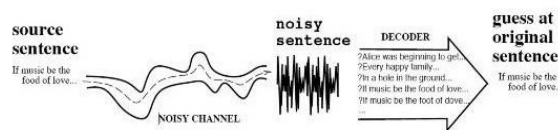
در مقاله حاضر، روشی ترکیبی بر مبنای گسترش و اصلاح پرس‌وجو، ارائه شده است. در این روش، ابتدا کلمه پرس‌وجو مورد ارزیابی قرار می‌گیرد. اگر این کلمه، جزو کلمات مستعد برای خطای بازشناسی، تشخیص داده شود، کلمات مشابه این پرس‌وجو نیز در نظر گرفته می‌شود. در بخش ادبیات موضوع با مفهوم مدل کانال نوفه‌ای، آشنا می‌شویم که آشنایی با این مفهوم، قبل از بیان روش پیشنهادی ضروری است.

۳- ادبیات موضوع

۳-۱- مدل کانال نویزی (NCM)

NCM چارچوبی است که در خطایاب املائی 2 ، پاسخ دهنده خودکار سوالات 3 ، بازشناسی گفتار 4 و ماشین ترجمه 5 استفاده شده است. در این مدل، هدف، یافتن کلمات صحیح جایگزین، بر اساس یک کلمه داده شده است. اگر NCM درست آموزش داده شود، با احتمال زیادی می‌تواند کلمه صحیح را حدس بزند [8]، [13]، [11]. از آن جا که یکی از مدل‌های استفاده‌شده در روش پیشنهادی، همین مدل است، این روش را با جزئیات بیشتر در این قسمت توضیح می‌دهیم.

تعریف: حروف الفبای Σ داده شده است، Σ^* را به‌عنوان یک مجموعه‌ای از تمامی رشته‌های محدود بر روی Σ در نظر بگیرید. فرض کنید فرهنگ لغت D زیر مجموعه‌ای از Σ^* باشد ($D \subset \Sigma^*$).



(شکل-۲): شمای کلی مدل کانال نوفه‌ای در اصلاح خطای جملات (Figure-2): Overview of Noisy Channel Model in Error correction of sentences

ماتریس کانال نوفه‌ای 6 ، ماتریسی به شکل زیر است:

$$\Gamma_{ws} = \Pr(s|w)$$

¹ Noisy Channel Model (NCM)

² Spell checkers

³ question answering

⁴ Speech recognition

⁵ Machine translation

⁶ Noisy channel

⁷ Decision function

⁸ Maximum Likelihood

⁹ Maximum a posteriori probability

¹⁰ Minimum edit distance

¹¹ Bayes' rule

۴-۱- الگوریتم روش پیشنهادی

ورودی: ۱) یک سری سند که متون موجود در آن‌ها از بازشناسی گفتار به دست آمده است. ۲) یک واژه‌نامه از واژگان زبان مورد نظر.

ا- ماتریس وقوع^{۲۲} مربوط به اسناد را به دست بیاور

ب- به ازای هر کلمه پرس‌وجو q

ب-۱- با استفاده از یک مجموعه داده آموزشی، $P_e(q)$ ، احتمال این‌راکه کلمه q دچار خطای بازشناسی بشود محاسبه می‌کنیم. توضیح بیشتر راجع به محاسبه $P_e(q)$ را در بخش ۵ (پیاده‌سازی و تحلیل نتایج) بیان خواهیم کرد.

ب-۲- کلمات مشابه با این کلمه و احتمال تبدیل هر یک از آن‌ها به کلمه پرس‌وجو را محاسبه کن.

ب-۳- احتمال‌های تبدیل محاسبه‌شده در مرحله قبل را مرتب کن.

ب-۴- بر مبنای یک مقدار آستانه کلمات مشابه دارای احتمال تبدیل بیشتر را به پرس‌وجو اضافه کن، وزن هر یک از کلمات پرس‌وجوی اضافه‌شده را متناسب با احتمال تبدیل محاسبه‌شده در مرحله قبل برای آن در نظر بگیر.

ب-۵- به هر یک از کلمات فعلی موجود در پرس‌وجو (اعم از کلمه اولیه و کلمات مشابه اضافه‌شده به پرس‌وجو)، وزنی اختصاص بده، وزن کلمه اولیه، بر مبنای نوع روش بازیابی تعیین می‌شود^{۲۳}. وزن کلمات مشابه متناسب با احتمال تبدیل محاسبه‌شده در مرحله ب-۱ است.

ج- روش بازیابی دلخواهی را روی اسناد اعمال کن. خروجی: اسناد مرتبط با کلمه پرس‌وجوی اصلی و کلمات مشابه آن.

در شکل (۳)، نمودار روش پیشنهادی آمده است. در ادامه، مراحل مختلف الگوریتم با تفصیل بیشتری، توضیح داده خواهند شد.

۴-۲- مفهوم ماتریس وقوع و نحوه محاسبه آن

در اسناد

ماتریس وقوع، ماتریسی است که سطر و ستون آن به ترتیب، واژگان و اسناد می‌باشند (این ترتیب می‌تواند معکوس هم باشد). اگر واژگان خاصی در یک سند مشخص

می‌دهیم اما کانال نوفه‌ای، کلمه s را به‌عنوان کلمه مشابه پرس‌وجو (که به‌احتمال دچار خطا شده است)، با احتمال $P(s|w)$ پیشنهاد می‌کند. در نتیجه، تابع تصمیم ما به‌صورت $\hat{w} = \operatorname{argmax}_w P(s|w)p(w)$ خواهد بود.

برای محاسبه \hat{w} ، روشی در [11] پیشنهاد شده است. در این روش، پارامترهای مدل، به‌دفعات تا رسیدن به وضعیت ایستا، تخمین زده می‌شوند.

در این قسمت، فقط مدل NCM اولیه توضیح داده شد. مدل‌های احتمالی کانال کامل‌تری نیز ارائه شده است. توتانوا و مور در [17] نشان دادند که ترکیب مدل NCM بر مبنای آوا و مدل NCM بر مبنای حرف، بازدهی بیشتری را نسبت به هرکدام از آن‌ها به‌تنهایی دارد. روش ترکیبی، تصحیح را با دقت ۹۵/۵۸٪ و سه گزینه بهتر با دقت ۹۹/۵۰٪ جواب را پیدا می‌کند [8].

۴- روش پیشنهادی

در این بخش، روش پیشنهادی توضیح داده می‌شود. این روش، بر مبنای گسترش پرس‌وجو است. به‌طور مشخص، پیش‌پردازشی به روش بازیابی اضافه کرده‌ایم که کلمات، احتمال وقوع خطا در آن‌ها، فهرست کلمات مشابه و احتمال تبدیل به آن‌ها را محاسبه می‌کند. در نهایت، کلمه صحیح، کلمه‌ای می‌شود که کمترین احتمال وقوع خطا را داشته باشد.

گسترش پرس‌وجو در روش‌ها و پژوهش‌های گذشته بر اساس گسترش اسناد مشابه صورت گرفته است؛ که یک فرایند زمان‌بر، ناکارآمد و پرهزینه به‌شمار می‌رود. حال آن‌که کارایی و سرعت از عوامل مهم در یک سامانه بازشناسی گفتار به‌شمار می‌رود. در روش پیشنهادی، علاوه‌بر بهبود در نتایج بازیابی، بهبود خوبی در سرعت بازیابی حاصل شده است. به این نحو که به‌جای گسترش اسناد بازشناسی، تنها پرس‌وجو گسترش داده می‌شود. با تعریف و تنظیم منطقی وزن کلمه اصلی در پرس‌وجو و کلمات مشابه بهبود خوبی در شناسایی خطاهای بازشناسی گفتار حاصل کردیم. این روش، مستقل از نوع روش بازیابی است؛ یعنی می‌توان آن را به‌عنوان یک پیش‌پردازش قبل از هر روش بازیابی قرار داد. در ادامه، ابتدا الگوریتم کلی روش بیان شده و سپس مراحل مختلف این الگوریتم به‌تفصیل توضیح داده می‌شود.

²² Occurrence matrix

²³ واضح است که در اکثر روش‌های بازیابی، وزنی به کلمه‌ی پرس‌وجو داده می‌شود.

نخست صحیح یا غلط بودن کلمه را مشخص کند و دوم این که در صورت غلط بودن معادل صحیح را حدس بزند. برای محاسبه احتمال تبدیل نیز روش زیر پیشنهاد می‌شود:

ا- به‌ازای کلمه‌ی پرس‌وجوی w ، تمام کلماتی را که فاصله‌ی لونشتاین آن‌ها با w از حد مشخصی کمتر است، به دست بیاور. به این کلمات، کلمات نامزد می‌گوییم.

فاصله‌ی لونشتاین، معیاری برای اندازه‌گیری تفاوت میان دو رشته است. این فاصله یکی از الگوریتم‌های مهم محاسبه فاصله ویرایشی به حساب می‌آید. فاصله لونشتاین میان دو رشته به‌صورت حداقل تعداد ویرایش‌های مورد نیاز برای انتقال از یک رشته به دیگری تعریف می‌شود. عملیات مجاز شامل «درج»، «حذف» و «جانشینی» تک‌نویسه است. مثال‌هایی از کلمات با فاصله لونشتاین برابر یک در زیر آمده است:

طبق ← طبقه (با درج «ه» به انتهای کلمه)

باز ← راز (با جانشینی «ر» به جای «ب»)

درونی ← درون (با حذف «ی» از انتهای کلمه)

برای آشنایی بیشتر با تئوری این روش به [16] و [6] مراجعه نمایید.

ب- به‌ازای هر کلمه نامزد c ،

ب-1- $p(w|c)$ - احتمال تبدیل کلمه نامزد c به w را محاسبه کن. این پارامتر، نشان‌دهنده‌ی احتمال تبدیل w به c است. در این مقاله، دو فرمول برای محاسبه چنین احتمالی ارائه شده است. لازم به ذکر است که این دو فرمول با هم معادل نیستند؛ بلکه هر یک به‌صورت شهودی، به‌نحوی نشان‌دهنده‌ی احتمال تبدیل یادشده هستند.

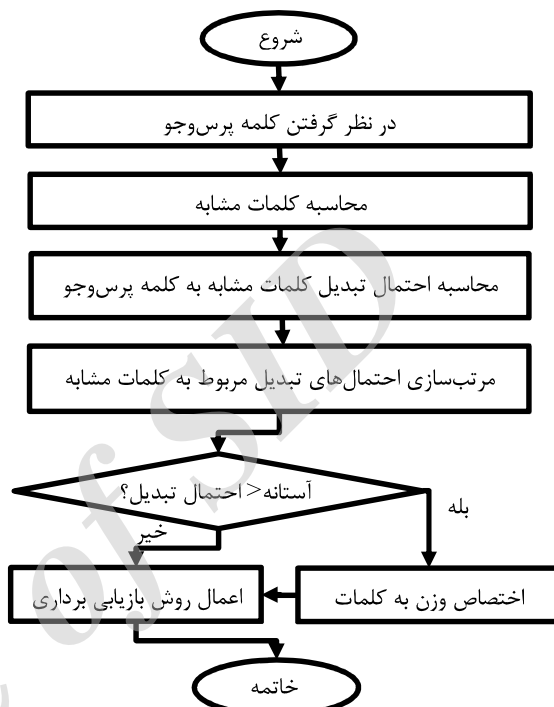
فرمول 1: ابتدا پارامتر نرخ تبدیل را به‌صورت زیر محاسبه و سپس با هنجارسازی آن، $p(w|c)$ را محاسبه می‌کنیم.

$$CR_{c,w} = \left(\frac{1}{l_{c,w}} \right) \times \left(\frac{L_w}{L_c} \right)^\alpha \quad (2)$$

که در این رابطه، $l_{c,w}$ ، فاصله‌ی لونشتاین در نظر گرفته‌شده بین c, w است. همچنین، L_c ، تعداد نویسه‌های کلمه نامزد، L_w ، تعداد نویسه‌های کلمه پرس‌وجو و α ، میزان تأثیر پارانتر دوم است.

در ادامه، هنجارسازی را نیز به‌صورت زیر انجام می‌دهیم:

وجود داشت، درایه مربوط به آن در ماتریس وقوع، یک و اگر واژگان در سند وجود نداشت این درایه، صفر خواهد شد. همچنین می‌توان به‌جای یک‌کردن درایه ماتریس در صورت وجود، مقدار درایه را برابر تعداد تکرار واژگان در سند مورد نظر دانست. توجه به این نکته حائز اهمیت است که برای کل اسناد ورودی، تنها یک ماتریس وقوع به‌دست آورده می‌شود.



(شکل-۳): نمودار جعبه‌ای الگوریتم پیشنهادی (Figure-3): Flowchart of the proposed algorithm

۳-۴- محاسبه کلمات مشابه و احتمال تبدیل آن‌ها به کلمه پرس‌وجو

اگر به‌ازای یک کلمه موجود در اسناد:

- بتوان تعیین کرد که آیا کلمه صحیح است (دچار خطای بازشناسی نشده است) یا خیر
- و اگر دچار خطا شده معادل صحیح کلمه چه می‌باشد؛

آن‌گاه به‌طور کامل تأثیر خطاهای بازشناسی از بین رفته است. واضح است که انجام این کار به شکل صددرصد امکان‌پذیر نیست؛ لیکن باید تلاش کرد تا حد امکان این کار انجام شود. هدف از این قسمت از الگوریتم، انجام چنین کاری است.

به‌طورطبیعی راه‌کار ارائه شده باید هر دو نکته اشاره‌شده در بالا را پوشش دهد؛ یعنی با دقت مناسبی

$ins[x,y]$ تعداد رخداد عمل درج حرف y بعد از حرف x است. به بیان دیگر تعداد دفعاتی که در اثر خطا، y بعد از x درج و کلمه صحیح به کلمه نادرست دیگری تبدیل شده است.

$sub[x,y]$ تعداد رخداد عمل جانشینی حرف y به جای حرف x است. به بیان دیگر تعداد دفعاتی که در اثر خطا، y جانشین حرف x شده و کلمه صحیح به کلمه نادرست دیگری تبدیل شده است.

$count[x]$ و $count[xy]$ به ترتیب نشان‌دهنده تعداد تکرار x و xy در مجموعه‌ی داده‌های آموزشی^{۲۶} مدل NCM است. ج- تا این جا، به‌ازای کلمه پرس‌وجو، تعدادی کلمه نامزد در نظر گرفته شده و برای هر یک، عددی به نام CR محاسبه شده است. اکنون مقادیر CR کلمات نامزد مربوط به پرس‌وجوی مورد نظر را به‌صورت نزولی مرتب کن.

د- بر مبنای یک مقدار آستانه، کلمات کاندید متناظر با CRهای بزرگ‌تر را به‌عنوان کلمات نامزد نهایی در نظر بگیر. این کلمات را در اصطلاح کلمات مشابه کلمه پرس‌وجو می‌نامیم.

ه- اکنون در هنگام جستجو واضح است که اگر خود کلمه پرس‌وجو عیناً یافت شود، انطباق حاصل است. همچنین اگر یکی از کلمات مشابه پرس‌وجو نیز در متن یافت شد، می‌گوییم در این موضع، خطای بازشناسی اتفاق افتاده و این کلمه مشابه را نیز انطباق در نظر می‌گیریم.

الگوریتم بالا، در واقع دو نکته ارائه شده در چند خط قبل را پوشش می‌دهد؛ یعنی هم موضع خطای بازشناسی را حدس می‌زند و هم کلمه جایگزین را پیشنهاد می‌دهد. هرچه فاصله لونشتاین بین کلمه پرس‌وجو و کلمه نامزد کمتر باشد، شباهت دو کلمه به یکدیگر بیشتر است. از طرف دیگر هرچه طول کلمه نامزد کوچک‌تر باشد، احتمال آن که پرس‌وجو بتواند جایگزین مناسب برای آن باشد، کمتر است؛ لذا پراتز دوم نیز در نظر گرفته شده است. در این رابطه، در بخش پیاده‌سازی، توضیحات دقیق‌تری خواهیم داد.

۴-۴- مرتب‌سازی احتمال‌های تبدیل مربوط به کلمات مشابه

در این مرحله، احتمالات تبدیل مربوط به کلمات مشابه را که در مرحله ب-۲ به دست آمده‌اند، به‌ترتیب نزولی مرتب

$$P(w|c) = \frac{CR_{c,w}}{\max(CR_{c,w})} \quad (۳)$$

به روش پیاده‌سازی شده توسط دو رابطه اخیر، در اصطلاح روش CR می‌گوییم.

فرمول ۲: در این رابطه، از مطالبی که در بحث مدل کانال نوفه‌ای بیان شد، استفاده می‌کنیم. در NCM رابطه زیر را در نظر گرفتیم.

$$P(w|c) = P(c|w) \times P(w) \quad (۴)$$

در این جا سعی می‌کنیم، برای محاسبه $P(w)$ و $P(c|w)$ در کاربرد خاص موضوع مقاله، روابطی را ارائه دهیم.

محاسبه $P(w)$ با استفاده از رابطه زیر انجام می‌شود.

$$P(w) = \frac{freq(w) + 0.5}{N} \quad (۵)$$

c ، کلمه نامزد است و w ، کلمه‌ای است که دچار خطای بازشناسی شده و قرار است، کلمه مشابه برای آن یافت شود.

برای محاسبه احتمال تبدیل، ابتدا باید احتمال کلمه w محاسبه شود. در ضمن $freq(w)$ تعداد رخداد کلمه w در مجموعه اسناد است. با توجه به روش باکس و تیائو می‌توان با استفاده از یک احتمال اولیه به یک توزیع پسین^{۲۴} برای یک واژه رسید. میزان در نظر گرفته شده برای این مقدار، تعداد رخداد به‌علاوه 0.5 است. این احتمال را تخمین احتمال مورد انتظار^{۲۵} می‌نامند [7]. احتمال شرطی $P(c|w)$ نیز به‌صورت زیر به‌دست می‌آید:

$$p(x|w) = \begin{cases} \frac{del[w_{i-1}, w_i]}{count[w_{i-1}, w_i]}, & \text{if deletion} \\ \frac{ins[w_{i-1}, x_i]}{count[w_{i-1}]}, & \text{if insertion} \\ \frac{sub[x_i, w_i]}{count[w_i]}, & \text{if substitution} \\ \frac{trans[w_i, w_{i+1}]}{count[w_i, w_{i+1}]}, & \text{if transposition} \end{cases} \quad (۶)$$

$del[x,y]$ ، تعداد رخداد عمل حذف حرف y بعد از x است. به بیان دیگر تعداد دفعاتی که در اثر خطا، y بعد از x حذف و کلمه صحیح به کلمه نادرست دیگری تبدیل شده است.

²⁶ training data set

²⁴ Posterior Distribution

²⁵ Expected Likelihood Estimate

$$\text{Weight}_w(c) = P(w|c) \times \text{Weight}(w) \times p_e(c) \quad (7)$$

در این معادله، $P(w|c)$ ، همان است که در رابطه (۴) تعریف شده است و $p_e(c)$ ، احتمال آن است که کلمه c دچار خطای بازشناسی بشود. $\text{weight}_w[c]$ ، وزن کلمه نامزد c به ازای کلمه پرس و جوی w است. $\text{weight}_w[c]$ ، وزنی است که روش بازیابی به کلمه پرس و جوی w می‌دهد.

۴-۷- اعمال بازیابی

در انتها، یک روش بازیابی دلخواه اعمال می‌شود. الگوریتم ارائه شده در این مقاله، هیچ محدودیتی روی نوع روش قرار نمی‌دهد و در واقع به مثابه یک پیش‌پردازش برای بازیابی به حساب می‌آید.

۵- پیاده‌سازی و تحلیل نتایج

در این قسمت از مقاله، ابتدا مجموعه داده مورد استفاده و محیط پیاده‌سازی توضیح داده می‌شود و سپس به بررسی و تحلیل کارایی روش پیشنهادی و مقایسه آن با روش‌های دیگر خواهیم پرداخت.

۵-۱- توصیف مجموعه داده

برای انجام این پژوهش نیاز به یک مجموعه از اسناد گفتاری همراه با متن اصلی اسناد و متن بازشناسی شده این اسناد بود. برای تبدیل اسناد گفتاری به متن معادل آن‌ها از مدل‌های آکوستیک مبتنی بر واج استفاده شده است [3]. مدل‌های آکوستیک مبتنی بر واج توسط دادگان فارسی‌دات بزرگ آموزش داده شدند. دادگان فارسی‌دات بزرگ توسط پژوهشگاه توسعه فناوری‌های پیشرفته خواجه نصیرالدین طوسی گردآوری شده و شامل ۱۴۰ ساعت گفتار از ۱۰۰ گوینده مختلف است. نیمی از این دادگان گفتاری (۷۰ ساعت) با استفاده از یک میکروفون دسک‌تاپ و بقیه با سه میکروفون دیگر ضبط شده است [1]. به منظور آموزش مدل‌های آکوستیک، داده آموزشی از سیگنال‌های ضبط شده با میکروفون دسک‌تاپ انتخاب شده است. داده آزمون نیز پاره‌گفتار ۲۷ از این دادگان است که از مجموعه آموزش جدا شده‌اند. مجموعه آزمون از هفت گوینده مختلف انتخاب شده و در مجموع حدود ۱,۵ ساعت است.

می‌کنیم. با توجه به واضح بودن این مرحله، درباره آن توضیح بیشتری نمی‌دهیم.

۴-۵- اضافه کردن کلمات مشابه به پرس و جو

در این مرحله بر مبنای یک مقدار آستانه تصمیم می‌گیریم که کدام یک از کلمات مشابه را همراه با پرس و جو در نظر بگیریم (معنای همراه بودن و تأثیر آن به زودی روشن می‌شود). بعد از در نظر گرفتن مقدار آستانه، کلمات مشابه با احتمال تبدیل بزرگ‌تر را همراه با پرس و جو در نظر می‌گیریم.

۴-۶- اختصاص وزن به کلمه پرس و جو و کلمات

مشابه

آنچه که تا به این مرحله انجام گرفته، بخش اعظمی از پیش‌پردازش قبل از اعمال روش بازیابی است. جهت تکمیل این پیش‌پردازش تنها یک نکته باقی مانده و آن تعیین وزن کلمات مشابه در گسترش پرس و جو است. این وزن در فرایند بازیابی نقش اساسی دارد و اگر وزن آن به صورت غیر منطقی تعیین شود، اثر پیش‌پردازش انجام گرفته تا این مرحله را خنثی می‌سازد. اگر وزن دهی کلمات مشابه را به خود بازیابی محول کنیم، روش بازیابی بدون توجه به احتمال تبدیل آن‌ها، وزنی را اختصاص می‌دهد که هیچ تأثیر در شناسایی خطاها نخواهد داشت و مثل این می‌ماند که ما به طور دقیق کلمات مشابه را به عنوان عبارت پرس و جو وارد کرده باشیم. سؤال اساسی آن است که چگونه این وزن را بر اساس پارامترهای پیش‌پردازش به گونه‌ای تنظیم کنیم که باعث بهبود در بازیابی خطاهای بازشناسی شود؟ جهت پاسخ به این سؤال باید رابطه احتمال تبدیل کلمه مشابه و احتمال خطای کلمه اصلی پرس و جو را با وزن کلمه پرس و جو، روشن سازیم.

وزن کلمات مشابه، باید متناسب با وزن کلمه پرس و جو باشد، زیرا در اصل، ما به دنبال کلمه پرس و جو هستیم نه کلمات مشابه؛ احتمال خطای کلمه اصلی پرس و جو نیز در وزن کلمات مشابه تأثیرگذار است. هرچه احتمال خطای کلمه پرس و جو بیشتر باشد، وزن کلمات مشابه باید به وزن کلمه اصلی پرس و جو نزدیک‌تر باشد. در بین کلمات مشابه نیز کلمه‌ای که احتمال تبدیل بیشتری را دارد، باید وزن بیشتری داشته باشد. بر اساس منطق بیان شده وزن کلمه مشابه در پرس و جو به صورت زیر تعریف می‌شود.

برنامه‌نویسی، نرم‌افزار Visual Studio 2013 است. برای ارزیابی کارایی از معیار اندازه F استفاده کرده‌ایم.

۵-۳- جزئیات پیاده‌سازی

یک روش خوب بازیابی خطای بازشناسی، باید در مقابل هر نوع پرس‌وجو به‌خوبی عمل کند؛ چون در پرس‌وجو، هم ممکن است، کلمه‌ای دچار خطا باشد و هم کلمه‌ای صحیح باشد. اگر کلمه‌ای دچار خطا باشد، روش، علاوه‌بر شناسایی کلمات صحیح، می‌بایست اسنادی را که دچار خطا شده‌اند، بازیابی کند و اگر کلمه صحیح باشد، روش باید حداقل به خوبی بازیابی معمولی عمل کرده و نتایج را بدتر نسازد. جهت کاهش خطاهای ناشی از بازشناسی گفتار، سه گام را باید طی نمود:

- گام نخست، تشخیص احتمال خطای کلمات است.
- گام دوم، یافتن کلمات مشابه، متناسب با کلمه‌ی مستعد خطا است. مشخص ساختن فهرستی از کلمات جایگزینی که ممکن است کلمات مستعد خطا، بعد از بروز خطا به آن‌ها تبدیل شوند (ما فهرستی از این کلمات را کلمات مشابه نامیده‌ایم). تنها یافتن کلمات مشابه کافی نیست، بلکه از آن‌جا که یک کلمه بعد از بروز خطا، با احتمال متفاوت به کلمات دیگر تبدیل می‌شود، محاسبه احتمال تبدیل اهمیت زیادی دارد.
- گام سوم، بازیابی اسنادی است که دچار خطای بازشناسی شده‌اند. این کار با اضافه کردن پارامترهای به‌دست‌آمده در دو گام قبلی، به یک روش بازیابی صورت می‌گیرد.

۵-۳-۱- محاسبه احتمال خطای کلمات پرس‌وجو

ما باید بتوانیم با احتمال خوبی تشخیص دهیم که کلمه پرس‌وجو تا چه اندازه مستعد خطای بازشناسی است. این گام از آن جهت اهمیت دارد که اگر روش، به‌اشتباه کلمات مشابه‌ای را به پرس‌وجویی که احتمال خطای بازشناسی آن کم است، اضافه کند، بازیابی در انتخاب اسناد، دچار خطا می‌شود و اثر منفی بر معیارهای بازیابی خواهد داشت. این احتمال از فرمول زیر محاسبه می‌شود:

$$P_e(q) = \frac{\text{count}_e(q)}{\text{count}(q)} \quad (8)$$

q، هر کلمه‌ای از مجموعه دادگان آموزشی می‌تواند باشد. نسبت تعداد دفعاتی که این کلمه دچار خطای بازشناسی شده به کل تکرارهای آن در دادگان، نشان‌دهنده احتمال خطای کلمه است. در پیش‌پردازش قبل از بازیابی به‌ازای تمام کلمات واژه‌نامه این احتمال محاسبه می‌شود.

مجموعه‌ی واژگان نیز شامل شصت‌هزار کلمه بوده است؛ در سامانه بازشناسی از مدل تری‌گرام^{۲۸} به‌عنوان مدل زبانی و به‌صورت اتصال نیمه [10] استفاده شده است. برای آموزش مدل تری‌گرام از پیکره متنی زبان فارسی (شامل حدود ۱۰۰ میلیون کلمه) استفاده شده است [5] و روش هموارسازی کاتز^{۲۹} بر روی آن اعمال شده است [12].

در واقع در این پژوهش از این مجموعه آزمون به‌عنوان مجموعه داده برای بازیابی اطلاعات گفتاری، استفاده شده است. این مجموعه داده شامل تعداد ۷۵۳ سند بوده و تعداد جملات هر یک از این اسناد بین ۱ تا ۱۲ جمله است. دقت بازشناسی کلمات این مجموعه در حدود ۸۰ درصد است. بعد از حذف کلمات توقف از میان کل اسناد، ۳۹۶ سند، شامل خطای بازشناسی هستند. اطلاعات مربوط به خطا، شامل شماره سندی که خطا در آن رخ داده، کلمه‌ای که دچار خطا شده و نیز کلمه‌ای که در اثر خطا به آن تبدیل شده است، در قالب یک فایل متنی از قبل موجود است.

جهت یافتن کلمات مشابه و محاسبه احتمال تبدیل آن‌ها، نیاز به یک واژگان اولیه است، این واژگان اولیه می‌تواند یک واژگان کامل از کلمات پرکاربرد فارسی باشد و یا تنها شامل کلمات کلیدی مؤثر در بازیابی باشد. در یک سامانه بازیابی اطلاعات، سایر کلمات که در بازیابی مؤثر نیستند، (مانند کلمات توقف) و به‌طورمعمول در تمامی اسناد تکرار می‌شوند، در نظر گرفته نمی‌شوند؛ زیرا تأثیر منفی بر آمار و اطلاعات استخراج‌شده از تحلیل خطاهای بازشناسی داشته و از دید بازیابی نیز کم‌ارزش هستند.

به‌منظور کارایی بهتر در محاسبات مربوط به بازیابی و معیارهای آن، از یک واژه‌نامه اولیه که نزدیک به ده‌هزار کلمه پر کاربرد فارسی را در خود دارد، استفاده کرده‌ایم.

۵-۲- محیط اجرا و ابزارهای استفاده‌شده

تمامی مراحل الگوریتم (روش‌های بازیابی- برداری، روش پیشنهادی، کشف خطاهای بازشناسی، روش NCM، محاسبه معیارهای بازیابی برداری) به‌صورت عملی کدنویسی شده‌اند و زبان برنامه‌نویسی مورد استفاده C# است.

محیط آزمون و اجرای روش، یک سامانه رایانه با CPU Core i5 به میزان حافظه ۴ گیگابایت است. برای اجرای سریع‌تر و بهره‌وری بیشتر، کدنویسی به‌نحوی انجام شده است که پردازش‌ها به‌صورت موازی صورت گیرد. محیط

²⁸ Trigram

²⁹ Katz

³⁰ F-measure

۵-۳-۲- یافتن کلمات مشابه

این بخش مهم‌ترین بخش روش پیشنهادی است. شناسایی دقیق و کامل کلمات مشابه امکان‌پذیر نیست؛ اما می‌توان با درصد خوبی کلمات مشابه را شناسایی کرد. برای یافتن کلمات مشابه که خروجی نهایی این مرحله است، باید از کلمات واژه‌نامه، کار را آغاز کرد و بر اساس معیارهای مشخصی، آن‌ها را محدود ساخت تا به کلمات مشابه برسیم، این کار یک فرایند سه مرحله‌ای است.

مرحله نخست: یافتن کلمات کاندید، از طریق فیلتر کردن کلمات واژه‌نامه بر اساس فاصله ویرایشی لونشتاین. مرحله دوم: محاسبه احتمال تبدیل کلمات نامزد.

مرحله سوم: انتخاب بهترین کلمات نامزد.

قبل از پرداختن به جزئیات هر یک از مراحل بالا، به معرفی معیاری جهت سنجش کارایی آن‌ها می‌پردازیم.

برای اندازه‌گیری و نمایش میزان اثربخشی روش‌ها در تشخیص کلمات نامزد، نیاز به یک معیار است. در این قسمت به دنبال یافتن یک چنین معیاری هستیم. بدین‌منظور معیاری به نام متوسط نرخ تشخیص (ADR) تعریف می‌کنیم (از این پس تا پایان پژوهش جهت سادگی و کوتاهی به آن نرخ تشخیص گفته می‌شود)

$$ADR = \frac{\sum_{e=1}^n |TWe \cap RWe|}{|TWe| + |FWe|} \quad (9)$$

که در آن e ، کلمه‌ای است که دچار خطای بازشناسی شده، RWe ، مجموعه کلمات مشابه واقعی برای کلمه e ام (کلمه‌ای که دچار خطای بازشناسی می‌شود، به‌اشتباه به کلمه یا کلمات مشابه دیگری تبدیل می‌شود که ما آن‌ها را کلمات مشابه واقعی می‌نامیم)، TWe ، مجموعه کلمات مشابه به‌دست‌آمده از طریق روش پیشنهادی برای کلمه e ام و n ، تعداد کل خطاهای بازشناسی از مجموعه داده‌های اسناد بازشناسی شده، می‌باشند. FWe نیز، مجموعه کلمات مشابه نادرست تشخیص داده‌شده، به‌زای کلمه e ام است.^{۳۴}

۵-۳-۳- یافتن کلمات نامزد

در این پژوهش، از الگوریتم لونشتاین برای یافتن کلمات نامزد استفاده شده است. به دو دلیل الگوریتم لونشتاین برای این کار مناسب تشخیص داده شد. دلیل نخست، وجود رابطه

^{۳۱} real related word (RW)

^{۳۲} true related word (TW)

^{۳۳} false related word (FW)

^{۳۴} با مطالعه مقالات و منابع مرتبط با موضوع، رابطه‌ای که بیانگر میزان اثربخشی روش‌ها باشد، مشاهده نشد؛ لذا روش ADR، ارائه شد که به‌طور شهودی، فرمول مناسبی به‌نظر می‌رسد.

معناداری بین تعداد خطاهای بازشناسی و فاصله لونشتاین است. همان‌طور که در قبل در بخش توصیف داده اشاره شد، ۳۹۶ سند، دارای خطای بازشناسی هستند. بعد از حذف کلمات توقف و تحلیل‌های صورت‌گرفته روی این اسناد، تعداد آن‌ها در جدول (۱) بیان شد.

دلیل دوم اطلاعاتی است که نشان می‌دهد، بیشتر خطاها ناشی از سه عملیات جانشینی، درج و حذف است که هر سه در محاسبه الگوریتم لونشتاین لحاظ می‌شوند. برای رسیدن به این واقعیت، توزیع خطاها را برای هر دو الگوریتم دامروا- لونشتاین و لونشتاین به‌دست آوردیم. نتایج نشان می‌دهد، انتقال دو حرف مجاور تأثیر زیادی در خطاهای بازشناسی ندارد؛ بنابراین فاصله لونشتاین، معیار مناسبی برای انتخاب کلمات نامزد، تشخیص داده شد و از آن‌جا که روی هم رفته ۷۲٪ خطاها در فاصله لونشتاین‌های ۱ و ۲ می‌باشند، توان خود را بر روی این بخش از خطاها متمرکز ساخته و کلمات نامزد را کلماتی از واژه‌نامه در نظر می‌گیریم که فاصله لونشتاین آن‌ها نسبت به کلمه خطا، کوچک‌تر یا مساوی ۲ باشد.

۵-۳-۴- محاسبه احتمال تبدیل کلمات نامزد

فاصله لونشتاین به‌عنوان یک فیلتر اولیه برای حذف کلمات نامزد نامرتب مناسب است، اما کافی نیست. در ابتدای امر، هر کلمه مستعد خطا ممکن است به هر یک از کلمات دیگر واژه‌نامه تبدیل شود؛ اما احتمال تبدیل آن‌ها متفاوت است؛ برای بسیاری از کلمات واژه‌نامه این احتمال صفر و تنها برای درصد اندکی از آن‌ها این احتمال بزرگتر از صفر است. هدف ما در این قسمت، محاسبه احتمال تبدیل برای کلمات نامزد به‌دست‌آمده از مرحله قبل است. بدیهی است که محاسبه دقیق این احتمال امکان‌پذیر نیست؛ اما با معیار ADR نشان خواهیم داد که احتمال تبدیل را تا حد معقولی درست محاسبه کرده‌ایم. برای محاسبه احتمال تبدیل، از دو روش استفاده کرده‌ایم و سپس آن‌ها را با یکدیگر مقایسه می‌کنیم.

۵-۳-۴-۱- مقایسه مدل‌ها بدون تعیین سطح آستانه

ابتدا اقدام به تعیین ضریب تأثیر α معرفی‌شده در معادله (۲) می‌کنیم و سپس مدل‌ها را با یکدیگر مقایسه می‌شوند. به‌زای مقادیر ۱۰، ۱۰۰، ...، α را بر معیار ADR محاسبه کرده، به‌نحوی که معیار ADR بیشینه شود. نتایج آن در شکل (۴) آمده است.

مدل CR به‌زای $\alpha=2$ بیشترین مقدار ADR را داراست.

(جدول-۱): تعداد خطاها بر حسب فاصله لونشتاین

(Table-1): The number of errors in terms of Loewenstein distance

فاصله لونشتاین	1	2	3	4	4	8	جمع کل خطاها
تعداد خطاها	181	103	79	26	6	1	396
بر حسب درصد	45.71%	26.01%	19.95%	6.57%	1.52%	0.25%	100%

(جدول-۲): نمونه‌ای از محاسبه احتمال تبدیل به‌ازای CR، برای کلمه "همکار"

(Table-2): An example of calculating the probability of conversion based on CR, for the word "Hamkar"

کلمه اصلی (w) = "همکار"		حداکثر فاصله لونشتاین = 2	
کلمه نامزد (x)	$CR_{x,w}$	$\max(CR_{x,w})$	$P_w(x)$
امکان	1.9%	6.4%	29.6875%
همراه	1.9%	6.4%	29.6875%
همکاری	6.4%	6.4%	100%
همواره	3.2%	6.4%	50%
...
مزار	0.3%	6.4%	14.0625%

(جدول-۳): نمونه‌ای از محاسبات مربوط به احتمال تبدیل

(Table-3): An example of calculating the probability of conversion

کلمه اصلی (w): "درونی"			حداکثر فاصله لونشتاین = 2			
کلمه نامزد (x)	Errors	$P(x w)$	$P(w)$	Distance	$P(x w)P(w)$	$P_w(x) \cdot 10^6$
روند	del[#s=>#^]-sub[د=>ی]-	3.62%E-3	3.31%E-2	2	1.19%E-6	0.119%
دینی	del[ر=>د]-sub[و=>ی]-	•	9.94%E-2	2	•	•
درون	del[ن=>ی]-	04.2017%	1.66%E-1	1	6.95%E-3	695.87%
دروس	del[و=>ن]-sub[س=>ی]-	•	3.31%E-2	2	•	•
دارویی	ins[د=>دا]-sub[ن=>ی]-	•	3.31%E-2	2	•	•
فروند	sub[د=>ف]-sub[د=>ی]-	•	3.31%E-2	2	•	•
کنونی	sub[د=>ک]-sub[ر=>ی]-	•	3.31%E-2	2	•	•
گروهی	sub[د=>گ]-sub[ن=>ی]-	1.74%E-4	9.94%E-2	2	1.72%E-7	0.0172%
ورودی	sub[د=>و]-sub[ن=>ی]-	•	3.31%E-2	2	•	•
درختی	sub[و=>خ]-sub[ن=>ی]-	2.28%E-4	3.31%E-2	2	7.54%E-8	7.54%E-3

رویکرد n-بهترین، نسبت به مدل دیگر بهتر عمل می‌کند و بهترین نتیجه در حالت ۱-بهترین است. هرچه n افزایش می‌یابد، ADR به دلیل افزایش کلمات نامزد نادرست، کمتر خواهد شد و این کاهش در هر دو مدل وجود دارد.

۳-۵-۲ مقایسه مدل‌ها بر اساس سطح آستانه

مطابق شکل ۶، نتایج ما نشان می‌دهد که مدل ارائه‌شده مبتنی بر NCM به‌ازای سطح آستانه بیشتر از ۰.۷، و مدل CR به‌ازای سطح آستانه ۰/۹، بیشترین ADR را دارند.

برای یافتن n-بهترین کلمات مشابه، هر یک از کلمات نامزد به‌دست‌آمده را بر اساس احتمال تبدیل، به‌صورت نزولی مرتب کرده و n تا از بهترین آن‌ها را به‌عنوان کلمات مشابه انتخاب می‌کنیم. برای یافتن n مناسب، هر دو مدل CR و NCM را به‌ازای مقادیر ۱۰، ۲، ۱ اجرا کردیم، که نتایج آن در شکل (۵) آمده است.

نمودار به وضوح نشان می‌دهد که مدل NCM در

^۱ علامت # به‌معنای نویسه null است، یعنی حرفی که وجود ندارد و برای نمایش حذف یک حرف از ابتدای کلمه یا درج یک حرف در ابتدای کلمه استفاده شده و بیشتر دلیل تکنیکی دارد.

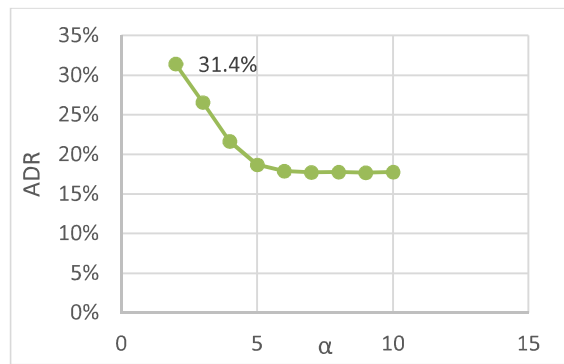
ارزیابی کارایی الگوریتم پیشنهادی، از یک روش بازیابی برداری با طرح وزن‌دهی استاندارد [15] Inc.ltc، بهره‌گرفتیم. وزن کلمه اصلی پرس‌وجو توسط روش بازیابی و وزن کلمات مشابه نیز طبق معادله (۷) مشخص می‌شود. برای ارزیابی کارایی روش ارائه‌شده، نتایج را یک بار با اضافه کردن پیش‌پردازش و یک بار بدون اضافه کردن آن به بازیابی عادی، براساس پرس‌وجوهای مختلف به دست آورده و با هم مقایسه می‌کنیم تا بهبود و کارایی الگوریتم پیشنهادی را نشان دهیم. بازیابی معمولی را الگوریتم اولیه^۱ و روش پیشنهادی را الگوریتم توسعه‌یافته^۲ می‌نامیم. باتوجه به آنکه روش ارائه‌شده، درواقع یک پیش‌پردازش به شمار می‌آید و همچنین به علت عدم وجود روشی مشابه که گسترش پرس‌وجو را بر مبنای ویژگی‌های موجود در بازشناسی گفتار (احتمال خطای بازشناسی) انجام دهد، ارزیابی کارایی به صورت توضیح داده‌شده، انجام گرفته است.

همان‌طور که در ابتدای این بخش اشاره کردیم، یک روش خوب بازیابی خطای بازشناسی، باید در مقابل هر نوع پرس‌وجو به خوبی عمل کند. برای ارزیابی کارایی الگوریتم توسعه‌یافته، از سه نوع پرس‌وجو استفاده کردیم.

- پرس‌وجوی بدون خطا: از کلماتی تشکیل یافته که دچار خطای بازشناسی نشده‌اند. انتظار می‌رود الگوریتم توسعه‌یافته در مقابل این‌گونه پرس‌وجوها بدتر از الگوریتم اولیه عمل نکند.
- پرس‌وجوی خطادار: پرس‌وجویی است که کلمات آن از بین کلماتی انتخاب می‌شوند که دچار خطای بازشناسی شده‌اند. انتظار می‌رود الگوریتم توسعه‌یافته در مقابل این‌گونه پرس‌وجوها بهتر از الگوریتم اولیه عمل کند.
- پرس‌وجوی تصادفی: کلمات این نوع پرس‌وجو به صورت تصادفی از واژگان انتخاب می‌شوند و ممکن است هر کدام از آن‌ها، کلمه خطا یا بدون خطا باشد. انتظار می‌رود الگوریتم توسعه‌یافته در مقابل این‌گونه پرس‌وجوها بهتر از الگوریتم اولیه عمل کند.

الگوریتم توسعه‌یافته با الگوریتم اولیه با هر سه نوع پرس‌وجو، مقایسه شده است. پرس‌وجوها، ۱۲۰ کلمه‌ای در نظر گرفته شده‌اند و نشان دادیم که الگوریتم توسعه‌یافته انتظارات را به خوبی برآورده می‌کند.

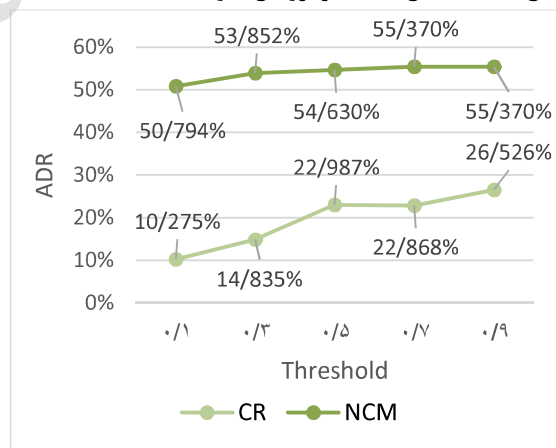
برای مقایسه نمونه‌های مختلف از نمودار F استفاده شد. برای این منظور، به تکرار محاسبات با تعداد مختلف



(شکل-۴): اندازه ADR بر حسب α

(Figure-4): ADR value in term of α

در پایان این بخش می‌توان نتیجه گرفت که مدل NCM، ۲۸٪ بهتر از مدل CR در تشخیص کلمات مشابه عمل می‌کند و این بهبودی در ۷۲٪ خطاها (مجموع خطاهایی که در فاصله لونشتاین ۱ و ۲ قرار دارند) است. پارامترهای مؤثر نیز در این بخش شناسایی شدند، نخستین پارامتر مؤثر فاصله لونشتاین حداکثر ۲ است که ۷۲٪ خطاها را پوشش می‌دهد. پارامتر دوم احتمال خطای تبدیل است که هرچه بیشتر باشد، احتمال تبدیل کلمه مستعد خطا به آن بیشتر است و پارامتر سوم، سطح آستانه‌ی ۰/۹ و ۰/۷ به ترتیب برای مدل CR و NCM می‌باشد. برای هر دو مدل انتظار می‌رود که معیارهای بازیابی نیز به همین میزان بهبود داشته باشند. نتایج حاصل از بازیابی که در قسمت بعدی بیان شده‌اند، این مطلب را روشن خواهند ساخت.



(شکل-۶): نمودار ADR و سطح آستانه برای دو مدل CR و

NCM

(Figure-6): ADR - Threshold chart for CR and NCM models

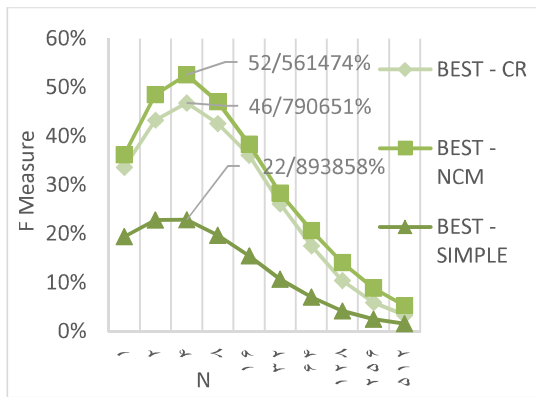
۴-۵- نتایج بازیابی

ایده اصلی بازیابی خطاهای بازشناسی در بخش قبل بیان شد. در این قسمت، فقط به ارائه نتایج بسنده می‌کنیم. تا این مرحله توانستیم پارامترهای مؤثر در شناسایی کلمات مشابه را برای هر دو مدل پیشنهادی شناسایی کنیم. جهت

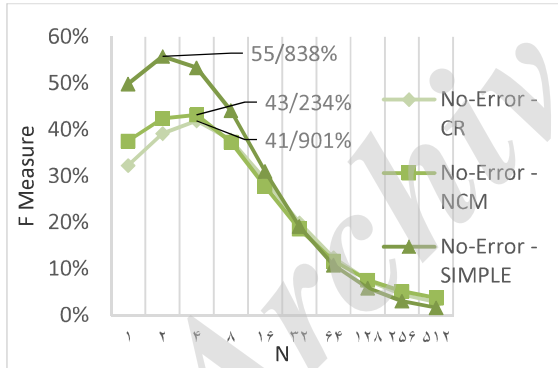
¹ Main algorithm

² Expanded algorithm

در شکل (۹) نتایج در مقابل پرس وجوهای بدون خطا در حالتی که از N به جای سطح آستانه استفاده شود، متفاوت است. الگوریتم توسعه یافته در مقابل این نوع پرس وجوها بدتر عمل می کند؛ اما در مرحله بعد وقتی از سطح آستانه استفاده می کنیم، نشان خواهیم داد که الگوریتم توسعه یافته در مقابل این نوع پرس وجوها نیز، با سطح آستانه بهتر عمل می کند. نکته دیگر آن که الگوریتم اولیه به ازای $N=2$ در مقابل پرس وجوهایی از این قبیل، خوب عمل می کند.



(شکل ۸): نمودار F-N در مقابل پرس وجوی خطا دار
(Figure-8): F - N chart for erroneous query



(شکل ۹): نمودار F-N در مقابل پرس وجوی بدون خطا
(Figure-9): F - N chart for correct query

۳-۴-۵- مقایسه نتایج اجرا با در نظر گرفتن سطح آستانه

در قسمت قبل برای مقایسه دو مدل، تعداد نتایج بازیابی (N) را تعیین می کردیم. در کاربردهای واقعی بازیابی اطلاعات، مدل ها باید به صورت خودکار، تعداد نتایج بازیابی شده را مشخص کنند. برای این منظور نیاز به تعریف سطح آستانه برای امتیاز اسناد، در الگوریتم بازیابی اطلاعات است. الگوریتم بازیابی از فاصله کسینوسی بین بردار اسناد و

اسناد بازیابی نیاز شده بود. در این پژوهش این عدد را با N نمایش می دهیم. به بیان دیگر عدد N نشان می دهد که الگوریتم بازیابی، چند سند را به عنوان اسناد مرتبط برگرداند. در تمامی مدل ها، بیشینه فاصله لونشتاین برای تعیین کلمات نامزد، ۲ در نظر گرفته شده است (دلیل آن در قبل در بخش ۳-۳-۵ توضیح داده شد).

۵-۴-۱- تعیین پارامترهای مدل

برای مقایسه دقیق و آسان، ابتدا پارامترهای هر یک از مدل ها را معرفی کرده و اعدادی را که قرار است، نتایج بر اساس آن ها مقایسه شوند مشخص می سازیم.

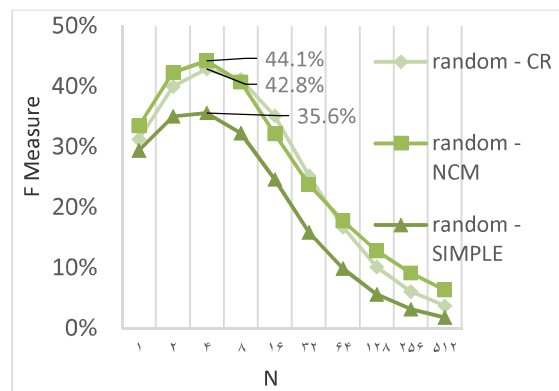
اعداد مناسب برای پارامترهای مؤثر در مدل NCM و CR، در قبل در بخش ۵-۳-۵ به دست آمدند. این پارامترها عبارتند از سطح آستانه α و بهترین کلمات نامزد که به ترتیب با T و M نمایش خواهیم داد. اعداد مناسب برای هر یک از آن ها $T=0.7$ و $M=1$ است.

پارامترهای مؤثر در مدل CR عبارتند از ضریب تأثیر و سطح آستانه که به ترتیب با α و T نشان داده شده اند و اعداد مناسب برای آن ها $\alpha=2$ و $T=0.9$ است.

۵-۴-۲- مقایسه نتایج اجرا، بدون در نظر گرفتن سطح آستانه

سه معیار «صحت^۱»، «فراخوانی^۲» و «معیار F » را برای این مدل ها با مقادیر N مختلف و پرس وجوهای مختلف، محاسبه کردیم. در شکل های ۷ و ۸، پارامتر اندازه F ، به ازای دو حالت پرس وجوی تصادفی و پرس وجوی خطا دار، نمایان است.

همان گونه که ملاحظه می شود، الگوریتم توسعه یافته در هر دو مدل (NCM و CR)، در مقابل پرس وجوهای تصادفی تا ۹٪ و در مواجهه با پرس وجوهای خطا دار تا ۳۰٪ بهبود را نسبت به بازیابی ساده نشان می دهد.



(شکل ۷): نمودار F-N در مقابل پرس وجوی تصادفی

(Figure-7); F - N chart for random query

¹ Precision
² recall

نشان می‌دهد. اطلاعات جدول نشان‌دهنده بهبود بر حسب درصد می‌باشد.

۶- نتیجه‌گیری و پیشنهادها

در این مقاله، یک پیش‌پردازش روی روش‌های بازیابی اطلاعات حاصل از متون بازنشاسی شده گفتاری، ارائه شد که به شکل ترکیبی از اصلاح و گسترش پرس‌وجو بود. ورودی‌های مسئله، اسناد متنی به‌دست‌آمده از بازنشاسی گفتار و پرس‌وجو، جهت یافتن اسناد مرتبط با کلمه پرس‌وجو بودند. در بخش ۵، نتایج پیاده‌سازی آورده شد و تحلیل‌های لازم روی آن صورت گرفت. دو مدل برای شناسایی کلمات مشابه و احتمال تبدیل آن‌ها معرفی شد.

برای سنجش کارایی مدل‌ها، معیار ADR تعریف و پارامترهای مؤثر در هر مدل شناسایی شد؛ سپس اندازه هر یک از پارامترها در جهت کارایی بهتر تعیین شد. روش ارائه‌شده به‌عنوان یک پیش‌پردازش قبل از بازیابی برداری قرار گرفت. در ضمن برای سنجش کارایی از معیارهای صحت، فراخوانی و F استفاده شد. برای مقایسه کارایی، یکبار روش بازیابی عادی و یکبار روش بازیابی همراه با پیش‌پردازش ارائه‌شده، به مجموعه داده اعمال شدند. مشخص شد الگوریتم توسعه‌یافته با سطح آستانه، بهتر از الگوریتم اولیه عمل می‌کند. همچنین مدل NCM عملکرد بهتری نسبت به مدل CR دارد. روش پیشنهادی با استفاده از مدل NCM، در مقابل خطاهای بازنشاسی، حدود ۳۰٪ بهبود در معیار F را ایجاد کرد. در آن، ترکیبی از روش بازیابی برداری، مدل NCM، مدل CR و الگوریتم لونشتاین، جهت شناسایی خطاهای حاصل از بازنشاسی به‌کار برده شد و به نتایج مثبتی رسیدیم.

ارائه نتایج بهتر و دقیق‌تر، نیازمند مجموعه داده وسیع‌تر و بیشتری در این زمینه است. هر چه داده‌های بیشتری برای آموزش مدل وجود داشته باشد، می‌توان نتایج دقیق‌تری را به‌دست آورد. کمبود داده‌های اولیه ناشی از خطای بازنشاسی فارسی، جهت آموزش بهتر مدل‌ها، از مشکلات پیش رو است. انتظار می‌رود که الگوریتم توسعه‌یافته، در بخش محاسبه احتمال خطای کلمات پرس‌وجو، با ترکیب مدل‌های زبانی و مدل‌های آکوستیکی، نتایج بهتری داشته باشد که پژوهش‌های بیشتری را در آینده می‌طلبد.

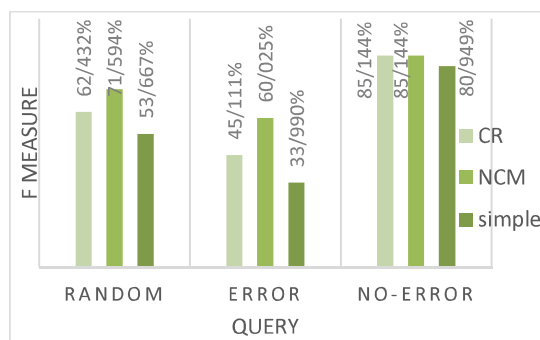
پرس‌وجو جهت رتبه‌بندی اسناد استفاده می‌کند. فاصله کسینوسی اسناد با پرس‌وجوی واردشده، بین ۰ تا ۱/۵۸ است. برای تفکیک اسناد مرتبط و نامرتب، یک سطح آستانه نیاز است. در این قسمت، سطح آستانه مناسب برای الگوریتم توسعه‌یافته و الگوریتم اولیه را به‌دست آورده و نتایج را بر اساس آن دوباره اجرا کرده و با یکدیگر مقایسه می‌کنیم.

برای محاسبه سطح آستانه، ابتدا رتبه‌های اسناد را به عددی با توزیع شبه‌نرمال تبدیل می‌کنیم و رتبه‌های جدید به‌دست آمده را به‌عنوان رتبه‌های اسناد در نظر می‌گیریم. فرمول تبدیل رتبه‌ها به توزیع شبه‌نرمال به‌صورت زیر است:

$$Rank_{new}(q, d) = \frac{Rank(q, d) - \mu}{\sigma} \quad (10)$$

که در این فرمول منظور از $Rank_{new}(q, d)$ ، عددی است که روش بازیابی در ازای پرس‌وجوی مشخص q ، به‌عنوان امتیاز یا رتبه برمی‌گرداند. μ ، میانگین و σ ، انحراف معیار رتبه‌های اسناد است. بدیهی است که میانگین اسناد به‌دست‌آمده برابر صفر و انحراف معیار آن‌ها برابر یک خواهد بود. از بین اسناد با رتبه‌های جدید، اسنادی را که رتبه آن‌ها از صفر بیشتر است، به‌عنوان اسناد مرتبط در نظر می‌گیریم.

بار دیگر نتایج را براساس سطح آستانه اجرا می‌کنیم. نتایج حاصل در شکل (۱۰) و جدول (۴) آمده است.



(شکل-۱۰): نمودار مقایسه معیار F براساس پرس‌وجوهای متفاوت

(Figure-10): F benchmark Comparison chart based on different queries

بر اساس اطلاعات این شکل و جدول، معیار F برای الگوریتم توسعه‌یافته، در برابر انواع پرس‌وجوها بهتر از الگوریتم اولیه عمل می‌کند. این در حالی است که بدون استفاده از سطح آستانه، الگوریتم اولیه در برابر پرس‌وجوی بدون خطا، بهتر از الگوریتم توسعه‌یافته عمل می‌کرد؛ اما با استفاده از سطح آستانه دیگر این‌گونه نیست. جدول (۵)، بهبود الگوریتم توسعه‌یافته را در معیارهای F، صحت و فراخوانی، بهتر

- [8] Brill, E., and Moore, R., "An Improved Error Model for Noisy Channel Spelling Correction", in 38th Annual meeting of Association for Computational Linguistics, Hong Kong, 2000.
- [9] Ghias, A., and et. al., "Query by Humming: Musical Information Retrieval in an Audio Database", in ACM Multimedia Conference, San Francisco, CA, USA, 1995, pp. 231-236.
- [10] Harper, M. P., and et. al., "Integrating Language Models with Speech Recognition", in AAAI94 Workshop on the Integration of Natural Language and Speech Processing, Seattle, Washington, USA, 1994, pp. 139-146.
- [11] Jurafsky, D., and Martin, J., Speech and Language Processing: An Introduction to Natural Language Processing, second ed. Prentice Hall, Pearson Education International, 2000.
- [12] Katz, S., "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer", IEEE Transactions on Acoustics, Speech and Signal Processing, vol 35(3), pp. 400-401, 1987
- [13] Kukich, K., "Techniques for Automatically Correcting Words in Text", ACM Computing Survey, vol 24, pp. 377-439, 1992.
- [14] Logan, B., and Van Thong, J., "Confusion-based query expansion for OOV words in spoken document retrieval", in 7th International Conference on Spoken Language Processing, Denver, Colorado, USA, 2002.
- [15] Manning, D. C., and et. al., An Introduction to Information Retrieval. Cambridge University Press, 2009.
- [16] Navarro, G., "A guided tour to approximate string matching", ACM Computing Surveys, vol 33(1), pp. 31-88, 2001.
- [17] Toutanova, K., Moore, R. C., "Pronunciation Modeling for Improved Spelling Correction", in 40th Annual meeting of Association for Computational Linguistics, Philadelphia, Pennsylvania, 2002, pp. 144-151.
- [18] Turunen, V., "Spoken Document Retrieval", in Department of Computer Science and Engineering Helsinki University of Technology, 2006.
- [19] Zhang, T., and Jay Kuo, C. C., "Content-based Classification and Retrieval of Audio", in 43th Annual Meeting-Conference on Advances Signal Processing, Algorithms, Architectures and Implementations, San Diego, 1998.

7-Reference

۷-مراجع

- [۱] شیخ زادگان، جواد، بیجن خان، محمود، "داده‌های گفتاری زبان فارسی"، در دومین کارگاه پژوهشی زبان فارسی و رایانه، دانشگاه تهران، ۱۳۸۵، صفحات ۲۴۷-۲۶۱.
- [1] J. Sheykhzadegan, M. Bijankhan, (2007). "Persian language speech data", in The Second Persian Language and Computer Research Workshop, University of Tehran, 2007, pp. 247-261.
- [۲] صرفجو، سعید، "چارچوبی جدید برای بازیابی اطلاعات به منظور استفاده در بازیابی صدای گفتاری فارسی"، پایان‌نامه منتشرشده کارشناسی ارشد، دانشکده فنی و مهندسی دانشگاه قم، قم، ایران، ۱۳۹۰.
- [2] Sarfjoo, S., "A New Framework for Information Retrieval to use in Persian Spoken Document Retrieval", Published master's dissertation, Faculty of Technical and Engineering of University of Qom, Qom, Iran, 2012.
- [۳] بحرانی، محمد، "به‌کارگیری ساختارهای وابسته به بافت در بازشناسی گفتار پیوسته مبتنی بر مدل مخفی مارکوف"، پایان‌نامه منتشرشده کارشناسی ارشد، دانشکده فنی و مهندسی دانشگاه صنعتی شریف، تهران، ایران، ۱۳۸۲.
- [3] Bahrani, M., "Using Context Dependent Structures in Continuous Speech Recognition based on Hidden Markov Model", Published master's dissertation, Faculty of Technical and Engineering of Sharif University of Technology, Tehran, Iran, 2004.
- [4] Abberley, D., and et. al., "THE THISL BROADCAST NEWS RETRIEVAL SYSTEM", in ESCA Tutorial and Research Workshop (ETRW) on Accessing Information in Spoken Audio, 1999, pp. 14-19.
- [5] Bijankhan, M., and et. al., "Lessons from Creation of a Persian Written Corpus: Peykare", Language Resources and Evaluation, vol. 45(2), pp. 143-164, 2011.
- [6] Black, Paul E., ed., "Levenshtein distance Dictionary of Algorithms and Data Structures", in National Institute of Standards and Technology, U. S., 14 August 2008. [online]. Available: www.nist.gov
- [7] Box, G., and Tiao, G., Bayesian Inference in Statistical Analysis. Massachusetts, Addison-Wesley, 1973.



روح الله دیانت تحصیلات کارشناسی

ارشد و دکترای خود را به ترتیب در سال‌های ۱۳۸۲ و ۱۳۸۹ در دانشگاه صنعتی شریف به اتمام رسانده و هم‌اکنون استادیار دانشگاه قم (گروه مهندسی کامپیوتر و فناوری اطلاعات) است. زمینه اصلی فعالیت وی، پردازش چندرسانه‌ای (صدا، تصویر و ویدئو) است. نشانی رایانامه ایشان عبارت است از: rouhollah.dianat@gmail.com



مرتضی علی احمدی تحصیلات

کارشناسی خود را در رشته علوم رایانه در دانشگاه قم، در سال ۱۳۹۱ و همچنین تحصیلات کارشناسی ارشد خود را در رشته مهندسی فناوری اطلاعات، در دانشگاه قم، در سال ۱۳۹۳ به پایان رسانده است. زمینه پژوهشی و مورد علاقه وی، پردازش صدا و تصویر است. نشانی رایانامه ایشان عبارت است از: morteza.ali.ahmadi@gmail.com



محمدیحیی اخلاقی تحصیلات

کارشناسی خود را در رشته علوم رایانه در دانشگاه قم به انجام رسانده است. همچنین تحصیلات کارشناسی ارشد را در رشته مهندسی فناوری اطلاعات، در دانشگاه قم در سال ۱۳۹۲ به اتمام رسانده است. او هم‌اکنون به‌عنوان مربی در مؤسسه آموزش عالی خاتم‌النبین کابل مشغول به کار است. نشانی رایانامه ایشان عبارت است از: yahya.akhlaghi@gmail.com



باقر باباعلی تحصیلات کارشناسی ارشد

و دکترای خود را در رشته هوش مصنوعی در دانشگاه صنعتی شریف و در سال‌های ۱۳۸۲ و ۱۳۸۹ به اتمام رسانده است. هم‌اکنون ایشان، عضو هیأت علمی گروه آمار و علوم رایانه پردیس علوم دانشگاه تهران هستند. زمینه‌های اصلی مورد علاقه ایشان، بازیابی اطلاعات و پردازش گفتار است. نشانی رایانامه ایشان عبارت است از: bagher.babaali@gmail.com