

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

روح الله دیانت<sup>۱</sup>، مرتضی علی احمدی<sup>۲</sup>، محمد یحیی اخلاقی<sup>۳</sup> و باقر باباعلی<sup>۴</sup>

<sup>۱</sup> گروه مهندسی فناوری اطلاعات، دانشکده فنی و مهندسی، دانشگاه قم، قم، ایران

<sup>۳</sup> گروه علوم کامپیوتر، دانشگاه خاتم التبیین، کابل، افغانستان

<sup>۴</sup> پردیس علوم، دانشکده ریاضی، آمار و علوم کامپیوتر، دانشگاه تهران، تهران، ایران



## چکیده

در این مقاله، یک پیش پردازش روی روش های بازیابی اطلاعات، ارائه می شود 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. ورودی های 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 خطای در بازشناسی است که ممکن است منجر به این شود که کلماتی که در واقع مرتبط هستند و به علت وقوع خطای بازشناسی دگرگون شده اند، مرتبط تشخیص داده نشوند.

ایده روش ارائه شده، تشخیص خطای بازشناسی در کلمات و درنظر گرفتن کلمات مشابه برای آن دسته از کلماتی است که به عنوان خطای تشخیص داده شده اند. برای تشخیص کلمه خطای، پارامتری به عنوان احتمال خطای در کلمه تعریف می شود که بزرگ بودن آن بیان گر امکان بیشتر وقوع خطای در کلمه است. همچنین برای تشخیص کلمات مشابه، ابتدا با استفاده از معیار فاصله لونشتاین، کلمات مشابه اولیه را پیدا می کنیم؛ سپس احتمال تبدیل این کلمات مشابه به کلمه پرس و جو اصلی، محاسبه می شود. کلمات مشابه معنایی، از بین کلماتی که احتمال تبدیل بیشتری دارند، بر اساس یک سطح آستانه انتخاب می شوند. اکنون در الگوریتم بازیابی، علاوه بر کلمه اصلی، کلمات مشابه آن نیز در جستجو، مرتبط در نظر گرفته می شوند. نتایج پیاده سازی ها نشان می دهد که الگوریتم ارائه شده، معیار F را به میزان حداقل ۳۰٪ بهبود می بخشد.

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

## Introducing a New information Retrieval Method Applicable for Speech Recognized Texts

Rouhollah Dianat<sup>1</sup>, MortezaAli Ahmadi<sup>2</sup> , MohammadYahya Akhlaghi<sup>3</sup> & Bagher Babaali<sup>4</sup>

<sup>1,2</sup> Department of Information Technology Engineering, Faculty of Technical and Engineering, University of Qom, Qom, Iran

<sup>3</sup> Lecturer in Department of Computer Science, University of Khatam Al Nabeyin, Kabol, Afghanistan

<sup>4</sup> 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 پرس‌وجوی داده شده را به ترتیب رتبه<sup>۶</sup>، به کاربر برمی‌گرداند.

در بازیابی صدا، پرس‌وجو می‌تواند به دو شکل کلی داده شود [۱۹]. در شکل نخست که به آن پرس‌وجو با نمونه گفته می‌شود، از یک نمونه صوتی به عنوان پرس‌وجو استفاده می‌شود. در این حالت هدف کاربر این است که پوشش‌های صوتی را که شامل نمونه صوتی ورودی یا نزدیک به آن هستند، پیدا کند. بسیاری از سامانه‌های بازیابی موسیقی و بازیابی گفتار آهنگین از این روش استفاده می‌کنند. در حالت خاصی از این روش، کاربر یک آهنگ یا موسیقی را زمزمه می‌کند و سامانه بازیابی، موسیقی یا آهنگ‌های مشابه با آن را از یک پایگاه داده یافته و فهرست آن‌ها را به کاربر بر می‌گرداند. به این روش پرس‌وجو با زمزمه<sup>۷</sup> می‌گویند [۹]. در شکل دوم که به آن پرس‌وجو با کلمه کلیدی<sup>۸</sup> گفته می‌شود، پرس‌وجو، یک نمونه متنی شامل یک یا چند کلمه کلیدی است. در این روش که بیشتر در بازیابی صدای گفتاری به کار می‌رود، هدف کاربر این است که پوشش‌های گفتاری را که شامل آن کلمات کلیدی یا در موضوعی نزدیک به آن‌ها هستند، پیدا کند. در حالتی که با صدای غیر گفتاری سروکار داریم، پرس‌وجوی متنی می‌تواند ویژگی‌ها یا توصیفی از صوت باشد که در قالب کلمات بیان می‌شود.

برای بازیابی صدای گفتاری به طور معمول از ترکیبی از فناوری‌های بازشناسی گفتار<sup>۹</sup> و بازیابی اطلاعات استفاده می‌شود؛ یعنی سندهای گفتاری ابتدا به متن معادل تبدیل شده و سپس عملیات بازیابی بر روی متن انجام می‌شود [۱۸]. در شکل زیر شمای کلی یک سامانه بازیابی صدای گفتاری نمایش داده شده است:

<sup>6</sup> Rank

<sup>7</sup> Query by humming

<sup>8</sup> Query by keyword

<sup>9</sup> Speech recognition

## ۱- مقدمه

اگر چه امروزه استفاده از سامانه‌های رایانه‌ای، جستجو و بازیابی اطلاعات<sup>۱</sup> را تسهیل کرده است؛ ولی جستجو و بازیابی، تاکنون به طور عمده در محتواهای متنی انجام می‌بذریفته است؛ در حالی که نیاز به استخراج اطلاعات از محتواهای غیر متنی مانند صدا نیز احساس می‌شود، هرچند این نیاز تا حدی با تولید فراداده<sup>۲</sup> برای محتواهای صوتی رفع می‌شود، ولی تولید فراداده کاری زمان بر و پرهیزینه است. بنابراین به دنبال روش یا روش‌هایی 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 روش‌ها، بازیابی مبتنی بر محتوا<sup>۳</sup> گفته می‌شود که بازیابی صدا یکی از این موارد محسوب می‌شود.

مسئله بازیابی صدا را می‌توان به صورت زیر تعریف کرد: یک مجموعه از مستندات صوتی<sup>۴</sup> در اختیار داریم؛ می‌خواهیم 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 پرس‌وجو<sup>۵</sup> نامیده می‌شود. با این تعریف، یک سامانه بازیابی صدا به طور کلی به این صورت عمل می‌کند که کاربر درخواست خود را به صورت پرس‌وجو به سامانه می‌دهد. سامانه بازیابی صدا با روش‌های خاص خود میزان شباهت و یا میزان ارتباط و نزدیکی پرس‌وجوی داده شده را، با تک‌تک سندهای صوتی موجود در بایگانی می‌سنجد و سپس فهرستی از شبیه‌ترین و

<sup>1</sup> Information retrieval

<sup>2</sup> Metadata

<sup>3</sup> Content base retrieval

<sup>4</sup> Audio documents

<sup>5</sup> Query



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 درنهایت، بخش ششم به نتیجه‌گیری و ارائه پیشنهادها برای انجام کارهای آینده اختصاص دارد.

## ۲-پیشینهٔ پژوهش

در مقدمهٔ مقاله حاضر، توضیحاتی در مورد مفهوم بازیابی صدای گفتاری ارائه و بیان شد که تمرکز مقاله حاضر، روی بازیابی صدای 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 راه کار کلی برای حل این مشکل، گسترش پرس‌وجو<sup>۱</sup> 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 متراffد بودن و یا ساختاری باشد. علاوه‌بر گسترش پرس‌وجو، روش‌های اصلاح پرس‌وجو نیز که به طور معمول به صورت 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 [۴] و در بعضی دیگر بر مبنای ویژگی‌های آکوستیکی (مانند [۱۴]) انجام گرفته است.

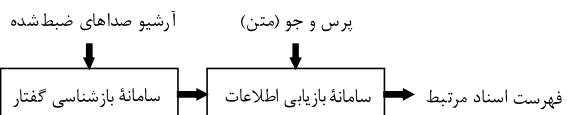
علاوه‌بر دو روش یادشده، روش‌های دیگری نیز ارائه شده است. به عنوان مثال، صرفجو در [۲]، روشی بر مبنای گسترش اسناد ارائه کرده است. در این روش، ابتدا کلمات تعدادی کلمه مشابه، استخراج می‌شوند. بدین ترتیب، در واقع به ازای هر سند، تعدادی سند مشابه به دست آمده است. به علت تعداد بسیار زیاد این اسناد مشابه، در مرحله بعد، تعداد محدودی از مناسب‌ترین آن‌ها انتخاب می‌شوند و

<sup>۱</sup>Query expansion

در این مقاله، تمرکز ما روی همین سامانه است، یک مشکل عمده این است که سامانه‌های بازشناسی گفتار همواره با درصدی خطا همراه هستند. درنتیجه ممکن است، سندی مرتبط با پرس‌وجو باشد، اما به علت خطای بازشناسی گفتار و درنتیجه وجود کلمات دچار اشکال در متن، به عنوان 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 به دست آمده 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 پیشینهٔ پژوهش و در بخش سوم

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

به عنوان مثال، برای حروف الفبای انگلیسی،  $\Sigma = \{a, b, c, \dots, y, z, A, B, \dots, Z, \dots\}$  آن  $s \subseteq \Sigma$ ، واژگانی از کلمات معتبر انگلیسی خواهد بود. اشتباهات و خطاهای مختلفی وجود دارد که شامل موارد زیر است.

- ۱- حذف ناخواسته حروف (برای مثال letter به جای letter)، اضافه شدن تصادفی حروف (برای مثال، mistake به جای mistake)
- ۲- جایه جایی حروف (برای مثال， received به جای received)
- ۳- جانشینی حروف (برای مثال finite به جای finite)

برای ساخت ماتریس کanal نوشه‌ای  $\Gamma$ ، 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  $s$  و  $w$  محاسبه شود.

هدف مدل کanal نوشه‌ای، یافتن کلمه صحیح از بین کلمات اشتباه ورودی است. برای انتخاب کلمه صحیح، نیاز به یک تابع تصمیم<sup>۷</sup> است. تابع تصمیم  $D \rightarrow \sigma: \Sigma^* \rightarrow \{0, 1\}$  تابعی است که کلمات اشتباه را گرفته و کلمه صحیح را بر می‌گرداند. روش‌های ساخت این تابع تصمیم شامل شیوه بیشترین احتمال (ML<sup>۸</sup>)، شوهی بیشترین احتمال پسین (MAP<sup>۹</sup>) و روش کمترین فاصله ویرایشی<sup>۱۰</sup> است.

در یک سیستم احتمالی، مامی خواهیم کلمه اشتباه  $s$  را اختیار داریم. مامی خواهیم از بین کلیه کلمات موجود در فرهنگ‌لغت، کلمه‌ای را انتخاب کنیم که موجب شود این احتمال، بیشینه شود. این کلمه را  $\hat{w}$  می‌نامیم. بر اساس قانون بیز<sup>۱۱</sup>، مامی رابطه زیر وجود دارد:

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

در مدل پیشنهادی، فرض شده است، کلمه‌ای را با احتمال  $p(w)$  به عنوان پرس‌وجو به سامانه بازیابی،

<sup>7</sup> Decision function

<sup>8</sup> Maximum Likelihood

<sup>9</sup> Maximum a posteriori probability

<sup>10</sup> Minimum edit distance

<sup>11</sup> Bayes' rule

درنهایت این اسناد، به همراه اسناد اصلی، در فرایند بازیابی، شرکت می‌کنند.

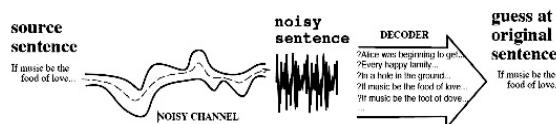
در مقاله حاضر، روشی ترکیبی بر مبنای گسترش و اصلاح پرس‌وجو، ارائه شده است. در این روش، ابتدا کلمه پرس‌وجو مورد ارزیابی قرار می‌گیرد. اگر این کلمه، جزو کلمات مستعد برای خطای بازشناسی، تشخیص داده شود، کلمات مشابه این پرس‌وجو نیز در نظر گرفته می‌شود. در بخش ادبیات موضوع با مفهوم مدل کanal نوشه‌ای، آشنا می‌شویم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

### ۳- ادبیات موضوع

#### ۱-۳- مدل کanal نوشه‌ای (NCM<sup>۱</sup>)

NCM چارچوبی است که در خطایاب املایی<sup>۲</sup>، پاسخ دهنده خودکار سوالات<sup>۳</sup>، بازشناسی گفتار<sup>۴</sup> و ماشین ترجمه<sup>۵</sup> 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]. از آن جا که یکی از مدل‌های استفاده شده در روش پیشنهادی، همین مدل است، این روش را با جزئیات بیشتر در این قسمت توضیح می‌دهیم.

تعريف: حروف الفبای  $\Sigma$  داده شده است،  $\sum^*$  را به عنوان یک مجموعه‌ای از تمامی رشته‌های محدود بر روی  $\Sigma$  در نظر بگیرید. فرض کنید فرهنگ‌لغت  $D$  زیر مجموعه‌ای از  $\sum^*$  باشد.  $(D \subset \Sigma^*)$ .



(شکل-۲): شماتیکی مدل کanal نوشه‌ای در اصلاح خطای جملات  
(Figure-2): Overview of Noisy Channel Model in Error correction of sentences

ماتریس کanal نوشه‌ای<sup>۶</sup>، ماتریسی به شکل زیر است:

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

<sup>1</sup> Noisy Channel Model (NCM)

<sup>2</sup> Spell checkers

<sup>3</sup> question answering

<sup>4</sup> Speech recognition

<sup>5</sup> Machine translation

<sup>6</sup> Noisy channel

## ۴-الگوریتم روش پیشنهادی

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

أَمَاتِرِيسْ وَقْوَعُ<sup>۱</sup> مُرْبُوطُ بِهِ اسْنَادِ رَا بِهِ دَسْتِ بِيَابُور  
ب- به ازای هر کلمه پرس‌وجو

ب-۱- با استفاده از یک مجموعه داده آموزشی، ( $P_e$ )<sup>(q)</sup><sub>e</sub>، احتمال این را که کلمه  $q$  دچار خطای بازشناسی بشود محاسبه می‌کنیم. توضیح بیشتر راجع به محاسبه ( $P_e$ )<sup>(q)</sup><sub>e</sub> 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 محاسبه شده 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 اضافه شده را متناسب با احتمال تبدیل محاسبه شده 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 اضافه شده به پرس‌وجو)، وزنی اختصاص بده، وزن کلمه اولیه، بر مبنای نوع روش بازیابی تعیین می‌شود.<sup>۲</sup> وزن کلمات مشابه متناسب با احتمال تبدیل محاسبه شده 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.

## ۲-۴-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

<sup>۱</sup> Occurrence matrix

<sup>۲</sup> واضح است که در اکثر روش‌های بازیابی، وزنی به کلمه پرس‌وجو داده می‌شود.

می‌دهیم اما کانال نویه‌ای، کلمه  $s$  را به عنوان کلمه مشابه پرس‌وجو (که به احتمال دچار خطای شده است)، با احتمال

$P(s|w)$  پیشنهاد می‌کند. درنتیجه، تابع تصمیم ما به صورت  $\hat{w} = \text{argmax}_w P(s|w)p(w)$  خواهد بود.

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

در این قسمت، فقط مدل NCM اولیه توضیح داده شد. مدل‌های احتمالی کانال کامل تری نیز ارائه شده است. توانای و مور در [17] نشان دادند که ترکیب مدل NCM بر مبنای آوا و مدل NCM بر مبنای حرف، بازدهی بیشتری را نسبت به هر کدام از آن‌ها به تنها 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. به طور مشخص، پیش‌پردازشی به روش بازیابی اضافه کرده‌ایم که کلمات، احتمال وقوع خطای در آن‌ها، فهرست کلمات مشابه و احتمال تبدیل به آن‌ها را محاسبه می‌کند. درنهایت، کلمه صحیح، کلمه‌ای می‌شود که کمترین احتمال وقوع خطای را داشته باشد.

گسترش پرس‌وجو در روش‌ها و پژوهش‌های گذشته بر اساس گسترش اسناد مشابه صورت گرفته است؛ که یک فاینده زمانی، ناکارآمد و پرهزینه به شمار می‌رود. حال آن که کارایی و سرعت از عوامل مهم در یک سامانه بازشناسی گفتار به شمار می‌رود. در روش پیشنهادی، علاوه‌بر بهبود در نتایج بازیابی، بهبود خوبی در سرعت بازیابی حاصل شده است. به این نحو که به جای 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 به عنوان یک پیش‌پردازش قبل از هر روش بازیابی قرار داد. در ادامه، ابتدا الگوریتم کلی روش بیان شده و سپس مراحل مختلف این الگوریتم به تفصیل توضیح داده می‌شود.

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

۱- بهازای کلمه‌ی پرس‌وجوی  $w$ ، تمام کلماتی را که فاصله لونشتاین آنها با  $w$  از حد مشخصی کمتر است، به دست بیاور. به این کلمات، کلمات نامزد می‌گوییم.  
فاصله 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 به صورت 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:

طبقه  $\leftarrow$  طبقه (با درج «ه» به انتهای کلمه)  
باز  $\leftarrow$  راز (با جانشینی «ر» به جای «ب»)  
درونى  $\leftarrow$  درون (با حذف «ى» از انتهای کلمه)

برای آشنایی بیشتر با تئوری این روش به [۱۶] و [۶] مراجعه نمایید.

ب- بهازای هر کلمه نامزد  $c$

ب-۱-  $P(w|c)$ ، احتمال تبدیل کلمه نامزد  $c$  به  $w$  را محاسبه کن. این پارامتر، نشان‌دهنده 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 به صورت شهودی، به نحوی نشان‌دهنده احتمال تبدیل یادشده هستند.

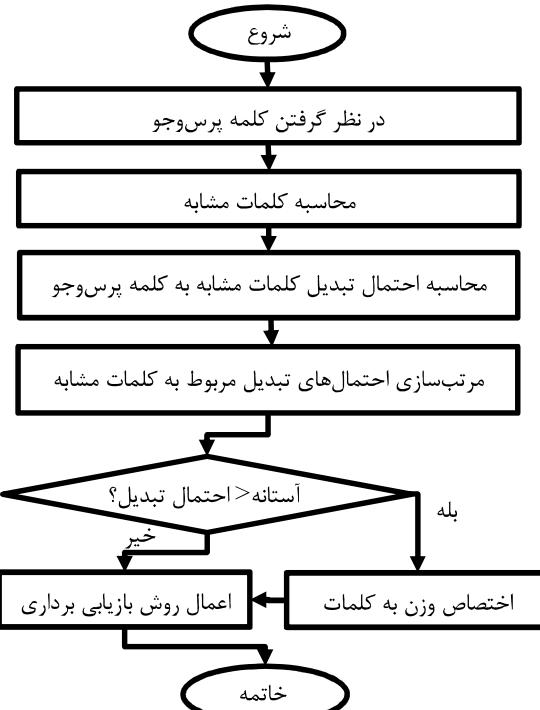
فرمول ۱: ابتدا پارامتر نرخ تبدیل را به صورت زیر محاسبه و سپس با هنجارسازی آن،  $P(w|c)$  را محاسبه می‌کنیم.

$$CR_{c,w} = \binom{1}{I_{c,w}} \times \binom{L_w}{L_c}^{\alpha} \quad (2)$$

که در این رابطه،  $I_{c,w}$ ، فاصله لونشتاین در نظر گرفته شده بین  $w$  و  $c$  است. همچنین،  $L_c$ ، تعداد نویسه‌های کلمه نامزد،  $L_w$ ، تعداد نویسه‌های کلمه پرس‌وجو و  $\alpha$ ، میزان تأثیر پرانتر دوم است.

در ادامه، هنجارسازی را نیز به صورت 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 به جای 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 به دست آورده می‌شود.



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

### ۴-۳-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 بازناسی نشده است) یا خیر
- و اگر دچار خطای شده معادل صحیح کلمه چه می‌باشد؛

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

به طور طبیعی راه کار ارائه شده باید هر دو نکته اشاره شده 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 خطأ،  $y$  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 خطأ،  $y$  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<sup>۳</sup> مدل NCM است. ج- تا این‌جا، به ازای کلمه پرس‌وجو، تعدادی کلمه نامزد در نظر گفته شده و برای هر یک، عددی به نام CR محاسبه شده است. اکنون مقادیر CR کلمات نامزد مربوط به پرس‌وجوی مورد نظر را به صورت نزولی مرتب کن.

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

- 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 بازناسنی اتفاق افتاده و این کلمه مشابه را نیز انطباق در نظر می‌گیریم.

الگوریتم بالا، درواقع دو نکته ارائه شده در چند خط قبل را پوشش می‌دهد؛ یعنی هم موضع خطای بازناسنی 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 پرانتر 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، به ترتیب نزولی مرتب

<sup>3</sup> training data set

$$P(w | c) = \frac{CR_{c,w}}{\max(CR_{cl,w})} \quad (3)$$

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

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

$$P(w | c) = P(c | w) \times P(w) \quad (4)$$

در این‌جا 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.

$$P(w) = \frac{freq(w) + 0.5}{N} \quad (5)$$

۵- کلمه نامزد است و  $w$ ، کلمه‌ای است که دچار خطای بازناسنی شده و قرار است، کلمه مشابه برای آن یافته 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 تیانو می‌توان با استفاده از یک احتمال اولیه به یک توزیع پسین<sup>۱</sup> برای یک واژه رسید. میزان در نظر گرفته شده برای این مقدار، تعداد رخداد به علاوه ۵/۰ است. این احتمال را تخمین احتمال مورد انتظار<sup>۲</sup> می‌نامند [7]. احتمال شرطی  $p(c | w)$

$p(x | w)$  نیز به صورت زیر به دست می‌آید:

$$\begin{cases} del[\mathcal{W}_{i-}, \mathcal{W}_i], & \text{if deletion} \\ count[\mathcal{W}_{i-}, \mathcal{W}_i], & \\ ins[\mathcal{W}_{i-}, \mathcal{X}_i], & \text{if insertion} \\ count[\mathcal{W}_{i-}], & \\ sub[\mathcal{X}_i, \mathcal{W}_i], & \text{if substitution} \\ count[\mathcal{W}_i], & \\ trans[\mathcal{W}_i, \mathcal{W}_{i+1}], & \text{if transposition} \\ count[\mathcal{W}_i, \mathcal{W}_{i+1}], & \end{cases} \quad (6)$$

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

<sup>1</sup> Posterior Distribution

<sup>2</sup> 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  $c$  چار خطای بازشناسی بشود.  $\text{weight}_w[c]$ ، وزن کلمه  $c$  نامزد  $c$  بهازای کلمه پرس‌وجوی  $w$  است.  $\text{weight}_w[c]$ ، وزنی است که روش بازیابی به کلمه پرس‌وجوی  $w$  می‌دهد.

#### ۷-۴-اعمال بازیابی

در انتهای، یک روش بازیابی دلخواه اعمال می‌شود. الگوریتم ارائه شده در این مقاله، هیچ محدودیتی روی نوع روش قرار نمی‌دهد و درواقع 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.

#### ۱-۵-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 [۳]. مدل‌های آکوستیک مبتنی بر واج توسط دادگان فارس‌داد بزرگ آموزش داده شدند. دادگان فارس‌داد 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 [۱]. بهمنظور آموزش مدل‌های آکوستیک، داده آموزشی از سیگنال‌های ضبط شده با میکروفون دسکتاپ انتخاب شده است. داده آزمون نیز پاره‌گفتار<sup>۱</sup> از این دادگان است که از مجموعه آموزش جدا شده‌اند. مجموعه‌ی آزمون از هفت گوینده مختلف انتخاب شده و درمجموع حدود ۱,۵ ساعت است.

<sup>۱</sup>Utterance

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

#### ۴-۵-اضافه کردن کلمات مشابه به پرس‌وجو

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

#### ۴-۶-اختصاص وزن به کلمه پرس‌وجو 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 پرس‌وجو 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 خطاهای نخواهد داشت و مثل این می‌ماند که ما به طور دقیق کلمات مشابه را به عنوان عبارت پرس‌وجو وارد کرده باشیم. سؤال اساسی آن است که چگونه این وزن را بر اساس پارامترهای پیش‌پردازش به‌گونه‌ای 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 پرس‌وجو را با وزن کلمه پرس‌وجو، روش سازیم.

وزن کلمات مشابه، باید متناسب با وزن کلمه پرس‌وجو باشد، زیرا در اصل، ما به دنبال کلمه پرس‌وجو هستیم نه کلمات مشابه؛ احتمال خطای کلمه اصلی پرس‌وجو نیز در وزن کلمات مشابه تأثیرگذار است. هرچه احتمال خطای کلمه پرس‌وجو بیشتر باشد، وزن کلمات مشابه باید به وزن کلمه اصلی پرس‌وجو 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 پرس‌وجو به صورت زیر تعریف می‌شود.



برنامه نویسی، نرم افزار Visual Studio 2013 است. برای ارزیابی کارایی از معیار اندازه F<sup>3</sup> استفاده کردایم.

### ۳-۳-جزئیات پیاده سازی

یک روش خوب بازیابی خطای بازشناسی، باید در مقابل هر نوع پرس و جو به خوبی عمل کند؛ چون در پرس و جو، هم ممکن است، کلمه‌ای دچار خطا باشد و هم کلمه‌ای صحیح باشد. اگر کلمه‌ای دچار خطا باشد، روش، علاوه بر 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 به دست آمده در دو گام قبلی، به یک روش بازیابی صورت می‌گیرد.

### ۴-۱-محاسبه احتمال خطای کلمات پرس و جو

ما باید بتوانیم با احتمال خوبی تشخیص دهیم که کلمه پرس و جو تا چه اندازه مستعد خطای بازشناسی است. این گام از آن جهت اهمیت دارد که اگر روش، به استبا کلمات مشابه‌ای را به پرس و جویی 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:

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

و، 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 به ازای تمام کلمات واژه‌نامه این احتمال محاسبه می‌شود.

<sup>3</sup> F-measure

مجموعه‌ی واژگان نیز شامل شصت‌هزار کلمه بوده است؛ در سامانه بازشناسی از مدل ترایگرم<sup>1</sup> به عنوان مدل زبانی و به صورت اتصال نیمه [10] استفاده شده است. برای آموزش مدل ترایگرم از پیکره متنی زبان فارسی (شامل حدود ۱۰۰ میلیون کلمه) استفاده شده است [5] و روش هموارسازی کاتز<sup>2</sup> بر روی آن اعمال شده است [12].

در واقع در این پژوهش از این مجموعه آزمون به عنوان 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. دقیق بازشناسی کلمات این مجموعه در حدود ۸۰ درصد است. بعد از حذف کلمات توقف از میان کل اسناد، سند، شامل خطای بازشناسی هستند. اطلاعات مربوط به خطای، 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 متنه از قبیل 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 پر کاربرد 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 به طور معمول در تمامی اسناد تکرار می‌شوند، در نظر گرفته نمی‌شوند؛ زیرا تأثیر منفی بر آمار و اطلاعات استخراج شده از تحلیل خطاهای بازشناسی داشته و از دید بازیابی نیز کم ارزش هستند.

به منظور 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 کردایم.

### ۴-۲-محیط اجرا و ابزارهای استفاده شده

تمامی مراحل الگوریتم (روش‌های بازیابی - برداری، روش پیشنهادی، کشف خطاهای بازشناسی، روش NCM، محاسبه معیارهای بازیابی برداری) به صورت عملی کدنویسی شده‌اند و زبان برنامه نویسی 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 به صورت موازی صورت گیرد. محیط

<sup>1</sup> Trigram

<sup>2</sup> Katz

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 خطاهای ناشی از سه عملیات جانشینی، درج و حذف است که هر سه در محاسبه الگوریتم لونشتاین لحاظ می‌شوند. برای رسیدن به این واقعیت، توزیع خطاهای را برای هر دو الگوریتم دامرو-لونشتاین و لونشتاین به دست 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٪ خطاهای در فاصله لونشتاین‌های ۱ و ۲ می‌باشد، توان خود را بر روی این بخش از خطاهای متمرکز ساخته و کلمات نامزد را کلماتی از واژه‌نامه در نظر می‌گیریم که فاصله لونشتاین آن‌ها نسبت به کلمه خط، کوچک‌تر یا مساوی ۲ باشد.

#### ۴-۳-۵-محاسبه احتمال تبدیل کلمات نامزد

فاصله لونشتاین به عنوان یک فیلتر اولیه برای حذف کلمات نامزد نامرتبط 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 به دست آمده 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 کردہ‌ایم و سپس آن‌ها را با یکدیگر مقایسه می‌کنیم. روش اول، محاسبه احتمال تبدیل با استفاده از CR می‌باشد. قبل از محاسبه احتمال تبدیل CR ابتدا فرمول (۲) را روشن‌تر می‌سازیم. این فرمول از دو قسمت تشکیل شده است. قسمت اول، معکوس فاصله لونشتاین کلمه کاندید با کلمه پرس‌وجو یا همان کلمه اولیه می‌باشد. دلیل آن هم با توجه به مطالعه بیان شده مشخص است. زیرا بیشتر کلماتی که دچار خطا می‌شوند به کلماتی با فاصله لونشتاین پایین تبدیل می‌شوند. لذا یک رابطه عکس بین

#### ۵-۲-۳-۵-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،  $RWe$ <sup>۱</sup>، مجموعه کلمات مشابه واقعی برای کلمه  $e$ ام (کلمه‌ای که دچار خطای بازشناسی می‌شود، به استباه به کلمه یا کلمات مشابه دیگری تبدیل می‌شود که ما آن‌ها را کلمات مشابه واقعی می‌نامیم)،  $TWe$ <sup>۲</sup>، مجموعه کلمات مشابه به دست آمده از طریق روش پیشنهادی برای کلمه  $e$ ام و  $n$ ، تعداد کل خطاهای بازشناسی از مجموعه داده‌های اسناد بازشناسی شده، می‌باشد.  $FWe$ <sup>۳</sup> نیز، مجموعه کلمات مشابه بازشناسی شده، می‌باشد. به ازای کلمه  $e$ ام است.<sup>۴</sup>.

#### ۵-۳-۵-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

<sup>1</sup> real related word (RW)

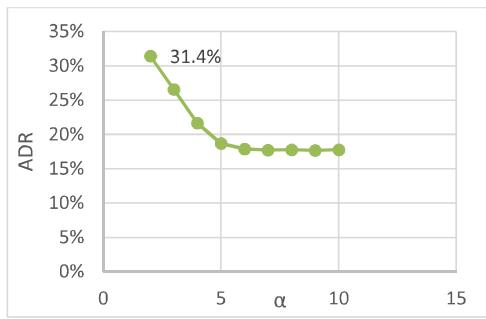
<sup>2</sup> true related word (TW)

<sup>3</sup> false related word (FW)

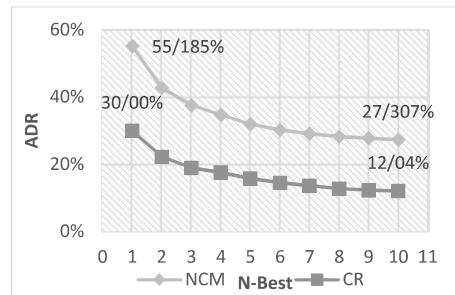
<sup>4</sup> با مطالعه مقالات و منابع مرتبط با موضوع، رابطه‌ای که بیانگر میزان اثربخشی روش‌ها باشد، مشاهده نشد؛ لذا روش ADR، ارائه شد که به طور شهودی، فرمول مناسبی به نظر می‌رسد.



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



(شکل-۴): اندازه ADR بر حسب  $\alpha$   
(Figure-4): ADR value in term of  $\alpha$



(شکل-۵): نمودار مقایسه مدل ها بر حسب ADR و N-BEST  
(Figure-5): Models comparison chart in terms of ADR and N-BEST

### ۲-۵-۳-۵- مقایسه مدل ها بر اساس سطح آستانه

طبق شکل(۶)، نتایج ما نشان می دهد که مدل ارائه شده مبتنی بر NCM به ازای سطح آستانه بیشتر از  $0.7$  و مدل به ازای سطح آستانه  $0.9$ ، بیشترین ADR را دارند.

در پایان این بخش می توان نتیجه گرفت که مدل  $n$ -BEST، NCM، بهتر از مدل CR در تشخیص کلمات مشابه عمل می کند و این بهبودی در  $72\%$  خطاهای (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 حداقل ۲ است که  $72\%$  خطاهای را پوشش می دهد. پارامتر دوم احتمال خطای تبدیل است که هرچه بیشتر باشد، احتمال تبدیل کلمه مستعد خطای آن بیشتر است و پارامتر سوم، سطح آستانه  $\alpha$  است که  $0.9$  بهترین مدل CR و NCM می باشد. برای هر دو مدل انتظار می رود که معیارهای بازیابی نیز به همین میزان بهبود داشته باشند. نتایج حاصل از بازیابی که در قسمت بعدی بیان شده اند، این مطلب را روشن خواهند ساخت.

احتمال تبدیل و فاصله لونشتاین وجود دارد. قسمت دوم نیز مربوط به رابطه طول کلمه پرس و جو نسبت به کلمه نامزد است. شواهد اولیه، رابطه مستقیم این نسبت با احتمال تبدیل را نشان می دهد. در مراحل بعد، این تأثیر را نشان خواهیم داد.

در جدول (۲)، نمونه ای از محاسبه احتمال تبدیل به ازای CR، برای کلمه "همکار" بیان شده است. روش دوم، محاسبه احتمال تبدیل از طریق معادله (۴) با استفاده از مدل NCM، است. در جدول (۳) نمونه ای از نتایج اجرا به ازای کلمه "دونو" آمده است.

بر اساس اطلاعات به دست آمده از جدول، کلمه "دونو"، نامزد مناسبی برای جایگزینی به نظر می رسد. در پایان این مرحله، توانستیم با استفاده از یک مدل تعریف شده جدید و با استفاده از مدل بهبود یافته NCM، یک مدل احتمال برای کلمات نامزد تعریف نماییم.

### ۵-۳-۵- انتخاب کلمات مشابه

در این مرحله، بر اساس احتمال بدست آمده در مرحله قبل، کلمات مشابه انتخاب خواهند شد. سپس میزان موفقیت هر یک از روش ها را در یافتن کلمات مشابه، توسط معیار ADR نشان خواهیم داد و نتایج آن ها را با یکدیگر مقایسه خواهیم کرد.

انتخاب کلمات مشابه، به دو روش انجام خواهد شد. یکی بر اساس سطح آستانه و دیگری بر اساس  $n$ -بهترین، به نحوی که معیار ADR بیشینه شود (معیار ADR در ۵-۲ توضیح داده شده است).

۱-۵-۳-۵- مقایسه مدل ها بدون تعیین سطح آستانه ابتدا اقدام به تعیین ضریب تأثیر  $\alpha$  معرفی شده در معادله (۲) می کنیم و سپس مدل ها را با یکدیگر مقایسه می شوند. به ازای مقادیر  $1, 0, \dots, 10$ ، تأثیر  $\alpha$  بر معیار ADR محاسبه کرده، به نحوی که معیار ADR بیشینه شود. نتایج آن در شکل (۴) آمده است.

مدل CR به ازای  $\alpha=2$  بیشترین مقدار ADR را دارد. برای یافتن  $n$ -بهترین کلمات مشابه، هر یک از کلمات نامزد به دست آمده را بر اساس احتمال تبدیل، به صورت نزولی مرتب کرده و  $n$  تا از بهترین آن ها را به عنوان کلمات مشابه انتخاب می کنیم. برای یافتن  $n$  مناسب، هر دو مدل CR و NCM را به ازای مقادیر  $1, 0, \dots, 10$  اجرا کردیم، که نتایج آن در شکل (۵) آمده است.

نمودار به وضوح نشان می دهد که مدل NCM در

<sup>۱</sup> n-best

(جدول-۱): تعداد خطاهای بر حسب فاصله لوئنستاین

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

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

(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"

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

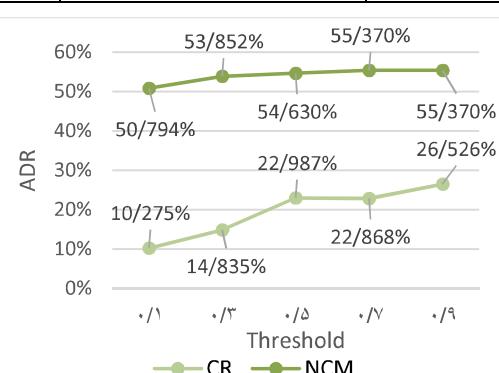
(جدول-۳): نمونه‌ای از محاسبات مربوط به احتمال تبدیل

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

حداکثر فاصله لوئنستاین=۲			کلمه اصلی(W): "دروند"				
$P_w(x) \cdot 10^6$	$P(x w)P(w)$	Distance	$P(w)$	$P(x w)$	Errors	کلمه نامزد(X)	
0.119%	1.19/E-6	2	3.31%E-2	3.62%E-3	$\text{del}[\#d=>\#] - \text{sub}[i=>d] -$	روند	
.	.	2	9.94%E-2	.	$\text{del}[d=>e] - \text{sub}[g=>i] -$	دینی	
95.87%	6.95/E-3	1	1.66%E-1	04.2017%	$\text{del}[n=>j] -$	درون	
.	.	2	3.31%E-2	.	$\text{del}[i=>w] - \text{sub}[o=>n] -$	دروس	
.	.	2	3.31%E-2	.	$\text{ins}[d=>a] - \text{sub}[n=>i] -$	دارویی	
.	.	2	3.31%E-2	.	$\text{sub}[d=>f] - \text{sub}[i=>d] -$	فروند	
.	.	2	3.31%E-2	.	$\text{sub}[d=>c] - \text{sub}[r=>n] -$	کنونی	
0.0172%	1.72/E-7	2	9.94%E-2	1.74%E-4	$\text{sub}[d=>g] - \text{sub}[n=>h] -$	گروهی	
.	.	2	3.31%E-2	.	$\text{sub}[d=>w] - \text{sub}[n=>d] -$	وروادی	
7.54%E-3	7.54/E-8	2	3.31%E-2	2.28%E-4	$\text{sub}[w=>n] - \text{sub}[x=>t] -$	درختی	

#### ۴-نتایج بازیابی

ایدها 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 ارزیابی کارایی الگوریتم پیشنهادی، از یک روش بازیابی برداری با طرح وزن دهنی استاندارد lnc.ltc [15]، بهره گرفتیم. وزن کلمه اصلی پرس و جو توسط روش بازیابی و وزن کلمات مشابه نیز طبق معادله (۷) مشخص می‌شود. برای ارزیابی کارایی روش ارائه شده، نتایج را یک بار با

(شکل-۶): نمودار ADR و سطح آستانه برای دو مدل CR و NCM  
(Figure-6): ADR - Threshold chart for CR and NCM models

<sup>۱</sup> علامت # به معنای 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 تکیه‌گذاری دارد.

الگوریتم بازیابی، چند سند را به عنوان اسناد مرتبط برگرداند. در تمامی مدل‌ها، بیشینه فاصله لونشتاین برای تعیین کلمات نامزد، ۲ در نظر گرفته شده است (دلیل آن در قبل در بخش ۳-۵ توضیح داده شد).

#### ۱-۴-۵-تعیین پارامترهای مدل

برای مقایسه دقیق و آسان، ابتدا پارامترهای هر یک از مدل‌ها را معرفی کرده و اعدادی را که قرار است، نتایج بر اساس آن‌ها مقایسه شوند مشخص می‌سازیم.

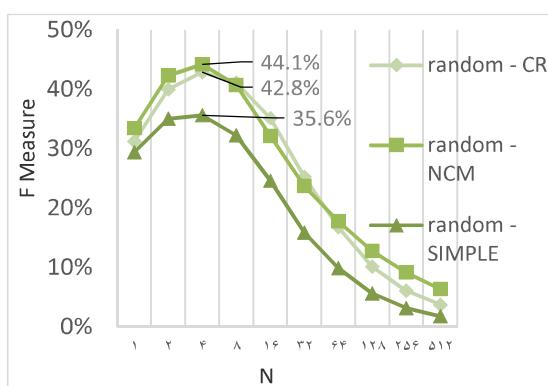
اعداد مناسب برای پارامترهای مؤثر در مدل NCM و CR، در قبل در بخش ۳-۵ به دست آمدند. این پارامترها عبارتند از سطح آستانه و  $n$ -بهترین کلمات نامزد که به ترتیب با  $T$  و  $M$  نمایش خواهیم داد. اعداد مناسب برای هر یک از آن‌ها  $T=7$  و  $M=1$  است.

پارامترهای مؤثر در مدل CR عبارتند از ضریب تأثیر و سطح آستانه که به ترتیب با  $\alpha$  و  $T$  نشان داده شده‌اند و اعداد مناسب برای آن‌ها  $\alpha=2$  و  $T=9$  است.

#### ۲-۴-۵-مقایسه نتایج اجرا، بدون درنظر گرفتن سطح آستانه

سه معیار «صحت<sup>۱</sup>»، «فراخوانی<sup>۲</sup>» و «معیار F» را برای این مدل‌ها با مقادیر  $N$  مختلف و پرس‌وجوهای مختلف، محاسبه کردیم. در شکل‌های ۷ و ۸، پارامتر اندازه  $F$ ، به ازای دو حالت پرس‌وجوی تصادفی و پرس‌وجوی خطدار، نمایان است.

همان‌گونه که ملاحظه می‌شود، الگوریتم توسعه‌یافته در هر دو مدل (CR و NCM)، در مقابل پرس‌وجوهای تصادفی تا ۹٪ و در مواجهه با پرس‌وجوهای خطدار تا ۳۰٪ بهبود را نسبت به بازیابی ساده نشان می‌دهد.



(شکل-۷): نمودار F-N در مقابل پرس‌وجوی تصادفی  
(Figure-7); F – N chart for random query

<sup>1</sup> Precision

<sup>2</sup> recall

اضافه کردن پیش‌پردازش و یک بار بدون اضافه کردن آن به بازیابی عادی، براساس پرس‌وجوهای مختلف بدست آورده و با هم مقایسه می‌کنیم تا بهبود و کارایی الگوریتم پیشنهادی را نشان دهیم. بازیابی معمولی را الگوریتم اولیه<sup>۱</sup> و روش پیشنهادی را الگوریتم توسعه‌یافته<sup>۲</sup> می‌نامیم. با توجه به آنکه روش ارائه شده، درواقع یک پیش‌پردازش به شمار می‌آید و همچنین به علت عدم وجود روشنی مشابه که گسترش پرس‌جو را بر مبنای ویژگی‌های موجود در بازناسی گفتار (احتمال خطای بازناسی) انجام دهد، ارزیابی کارایی به صورت توضیح داده شده، انجام گرفته است.

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

- پرس‌وجوی بدون خطای بازناسی تشکیل یافته که دچار خطای بازناسی نشده‌اند. انتظار می‌رود الگوریتم توسعه‌یافته در مقابل این‌گونه پرس‌وجوهای بدتر از الگوریتم اولیه عمل نکند.

- پرس‌وجوی خطدار: پرس‌وجویی است که کلمات آن از بین کلماتی انتخاب می‌شوند که دچار خطای بازناسی نشده‌اند. انتظار می‌رود الگوریتم توسعه‌یافته در مقابل این‌گونه پرس‌وجوهای بهتر از الگوریتم اولیه عمل کند.

- پرس‌وجوی تصادفی: کلمات این نوع پرس‌جوی به صورت تصادفی از واژگان انتخاب می‌شوند و ممکن است هر کدام از آن‌ها، کلمه خطای یا بدون خطای باشد. انتظار می‌رود الگوریتم توسعه‌یافته در مقابل این‌گونه پرس‌وجوهای بهتر از الگوریتم اولیه عمل کند.

الگوریتم توسعه‌یافته با الگوریتم اولیه با هر سه نوع پرس‌جو، مقایسه شده است. پرس‌وجوهای ۱۲۰ کلمه‌ای درنظر 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 اسناد بازیابی نیاز شده بود. در این پژوهش این عدد را با  $N$  نمایش می‌دهیم. به بیان دیگر عدد  $N$  نشان می‌دهد که

<sup>1</sup> Main algorithm

<sup>2</sup> Expanded algorithm

پرس‌وجو جهت رتبه‌بندی اسناد استفاده می‌کند. فاصله کسینوسی اسناد با پرس‌وجوی واردشده، بین ۰ تا ۱/۵۸ است. برای تفکیک اسناد مرتبط و نامرتبط، یک سطح آستانه نیاز است. در این قسمت، سطح آستانه مناسب برای الگوریتم توسعه‌یافته و الگوریتم اولیه را به دست آورده و نتایج را بر اساس آن دوباره اجرا کرده و با یکدیگر مقایسه می‌کنیم. برای محاسبه سطح آستانه، ابتدا رتبه‌های اسناد را به اعدادی با توزیع شبهنرمال تبدیل می‌کنیم و رتبه‌های جدید به دست آمده را به عنوان رتبه‌های اسناد در نظر می‌گیریم. فرمول تبدیل رتبه‌ها به توزیع شبهنرمال به صورت 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  $\mu$ ، به عنوان امتیاز یا رتبه برمی‌گرداند.  $\mu$ ، میانگین و  $\sigma$ ، انحراف معیار رتبه‌های اسناد است. بدیهی است که میانگین اسناد به دست آمده 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، به عنوان اسناد مرتبط در نظر می‌گیریم.

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

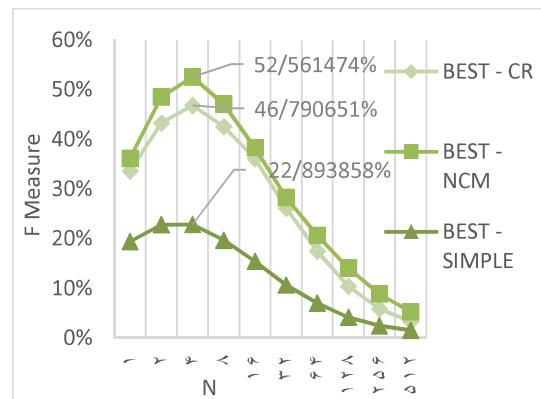


(شکل-۱۰): نمودار مقایسه معیار F براساس پرس‌وجوهای متفاوت

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

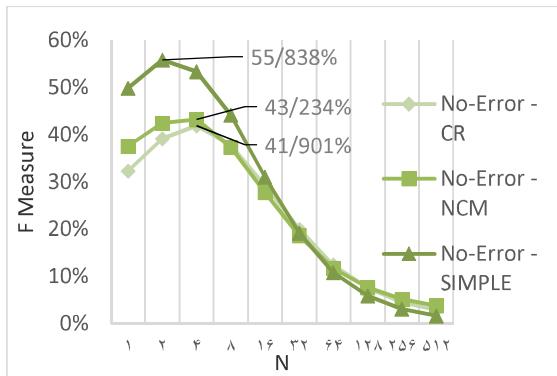
بر اساس اطلاعات این شکل و جدول، معیار F برای الگوریتم توسعه‌یافته، در برابر انواع پرس‌وجوها بهتر از الگوریتم اولیه عمل می‌کند. این در حالی است که بدون استفاده از سطح آستانه، الگوریتم اولیه در برابر پرس‌وجوی بدون خطای، بهتر از الگوریتم توسعه‌یافته عمل می‌کرد؛ اما با استفاده از سطح آستانه دیگر این گونه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

در شکل (۹) نتایج در مقابل پرس‌وجوهای بدون خطای در حالتی که از N به جای سطح آستانه استفاده شود، متفاوت است. الگوریتم توسعه‌یافته در مقابل این نوع پرس‌وجوها بدتر عمل می‌کند؛ اما در مرحله بعد وقتی از سطح آستانه استفاده می‌کنیم، نشان خواهیم داد که الگوریتم توسعه‌یافته در مقابل این نوع پرس‌وجوها نیز، با سطح آستانه بهتر عمل می‌کند. نکته دیگر آن که الگوریتم اولیه به ازای N=۲ در مقابل پرس‌وجوهایی از این قبیل، خوب عمل می‌کند.



(شکل-۸): نمودار F-N در مقابل پرس‌وجوی خطای

(Figure-8): F – N chart for erroneous query



(شکل-۹): نمودار F-N در مقابل پرس‌وجوی بدون خطای

(Figure-9): F – N chart for correct query

### ۴-۳-۴-۵- مقایسه نتایج اجرا با درنظر گرفتن سطح آستانه

در قسمت قبل برای مقایسه دو مدل، تعداد نتایج بازیابی (N) را تعیین می‌کردیم. در کاربردهای واقعی بازیابی اطلاعات، مدل‌ها 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

## 7-Reference

## 7-مراجع

- [1] شیخ زادگان، جواد، بیجن خان، محمود، "داده های 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 Langugae and Computer Research Workshop, University of Tehran, 2007, pp. 247-261.
- [2] صرفجو، سعید، "چارچوبی جدید برای بازبایی اطلاعات به منظور استفاده در بازبایی صدای گفتاری فارسی"، پایان نامه منتشر شده 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.
- [3] بحرانی، محمد، "به کارگیری ساختارهای وابسته به بافت در بازناسی گفتار پیوسته مبتنی بر مدل مخفی مارکوف"، پایان نامه منتشر شده 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.

نشان می دهد. اطلاعات جدول نشان دهنده بهبود بر حسب درصد می باشد.

## ۶-نتیجه گیری و پیشنهادها

در این مقاله، یک پیش پردازش روی روش های بازبایی اطلاعات حاصل از متون بازناسی شده گفتاری، ارائه شد که به شکل ترکیبی از اصلاح و گسترش پرس و جو بود. ورودی های مسئله، استاد متنی به دست آمده از بازناسی گفتار و پرس و جو، جهت یافتن اسناد مرتبط با کلمه پرس و جو بودند. در بخش ۵، نتایج پیاده سازی آورده شد و تحلیل های لازم روی آن صورت گرفت. دو مدل برای شناسایی کلمات مشابه و احتمال تبدیل آن ها معرفی شد. برای سنجش کارایی مدل ها، معیار ADR تعریف و پارامترهای مؤثر در هر مدل شناسایی شد؛ سپس اندازه هر یک از پارامترها در جهت کارایی بهتر تعیین شد. روش ارائه شده به عنوان یک پیش پردازش قبل از بازبایی برداری قرار گرفت. در ضمن برای سنجش کارایی از معیارهای صحت، فراخوانی و F استفاده شد. برای مقایسه کارایی، یکبار روش بازبایی عادی و یکبار روش بازبایی همراه با پیش پردازش ارائه شده، به مجموعه داده اعمال شدند. مشخص شدن الگوریتم توسعه یافته با سطح آستانه، بهتر از الگوریتم اولیه عمل می کند. همچنین مدل NCM عملکرد بهتری نسبت به مدل CR دارد. روش پیشنهادی با استفاده از مدل NCM، در مقابل خطاهای بازناسی، حدود ۳۰٪ بهبود در معیار F را ایجاد کرد. در آن، ترکیبی از روش بازبایی برداری، مدل NCM، مدل CR و الگوریتم لونشتاین، جهت شناسایی خطاهای حاصل از بازناسی به کار برده شد و به نتایج مثبتی رسیدیم.

ارائه نتایج بهتر و دقیق تر، نیازمند مجموعه داده وسیع تر و بیشتری در این زمینه است. هر چه داده های بیشتری برای آموزش مدل وجود داشته باشد، می توان نتایج دقیق تری را به دست آورد. کمبود داده های اولیه ناشی از خطای بازناسی فارسی، جهت آموزش بهتر مدل ها، از مشکلات پیش رو است. انتظار می رود که الگوریتم توسعه یافته، در بخش محاسبه احتمال خطای کلمات پرس و جو، با ترکیب مدل های زبانی و مدل های آکوستیکی، نتایج بهتری داشته باشد که پژوهش های بیشتری را در آینده می طلبد.



**روح الله دیانت تحصیلات کارشناسی ارشد و دکترای خود را به ترتیب در سال های 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 رایانمه 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 رایانمه 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 هم اکنون به عنوان مری در مؤسسه آموزش عالی خاتم النبیین کابل مشغول به کار است.**

**نشانی رایانمه ایشان عبارت است از:**

**yahya.akhlaghi@gmail.com**



**باقر بابعلی تحصیلات کارشناسی ارشد و دکترای خود را در رشته هوش مصنوعی در دانشگاه صنعتی شریف و در سال های ۱۳۸۲ و ۱۳۸۹ به اتمام رسانده است. هم اکنون ایشان، عضو هیأت علمی گروه آمار و علوم رایانه پردیس علوم دانشگاه تهران هستند. زمینه های اصلی مورد علاقه ایشان، بازیابی اطلاعات و پردازش گفتار است.**

**نشانی رایانمه ایشان عبارت است از:**

**bagher.bababli@gmail.com**

[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 43rd Annual Meeting-Conference on Advances Signal Processing, Algorithms, Architectures and Implementations, San Diego, 1998.

