تشخیص دست‌نوشتهٔ برخط فارسی با استفاده از مدل زبانی و کاهش قوانین نگارش کاربر

سلمان مسکنتی[1] و احمد کشاورز[2*]

[1] آزمایشگاه ماشین بینایی و هوش مصنوعی، دانشگاه خلیج فارس، بوشهر، ایران

[2] گروه مهندسی برق، دانشکده مهندسی، دانشگاه خلیج فارس، بوشهر، ایران

چکیده

پیوسته‌بودن کلمات فارسی و وجود تنوع بسیار زیاد رسم‌الخط این زبان و همچنین شکل‌های متنوع حروف فارسی بسته به محل قرارگیری‌شان در کلمه، تشخیص دست‌نوشته‌های فارسی را به چالش کشانده‌اند. مهم‌ترین اشکال در اغلب روش‌های بازشناسی بی‌توجهی به بافت جمله است که باعث می‌شود در مواردی که کلمهٔ ورودی اشتباه بازشناسی می‌شود، واژه‌ای با ظاهر درست در جمله‌ای نابه‌جا به کار رود. طراحی مدلی که بتواند بافت جمله را به‌خوبی تحلیل کند، مستلزم دراختیارداشتن منابع زبانی حجیمی است که نمایندهٔ خوبی از زبان مورد بازشناسی باشند. در این مقاله روش جدیدی برای بازشناسی کلمات برخط فارسی ارائه شده است که با استفاده از بافت جمله سعی در بهبود بازشناسی دارد. فرآیند بازشناسی معرفی‌شده در این نوشتار به این صورت است که ابتدا علائم و بدنهٔ زیرکلمات دست‌نوشتهٔ ورودی تفکیک شده و بدنهٔ هر زیرکلمه و علائم آن مشخص می‌شود؛ سپس علائم زیرکلمات تشخیص داده‌شده و بر اساس آن مجموعه‌ای از واژگان به‌عنوان فرضیه در نظر گرفته می‌شوند؛ به هر فرضیه بر اساس میزان شباهت آن به دست‌نوشته ورودی امتیازی تعلق می‌گیرد و بر اساس امتیاز حاصله محتمل‌ترین فرضیات مشخص می‌شوند. سپس این رویه توسط مدل زبانی برای یافتن فرضیات محتمل‌تر، هدایت می‌شود. نتایج آزمایش‌های به‌عمل‌آمده نشان می‌دهد که کاهش قابل توجهی در نرخ خطای بازشناسی کلمات حاصل شده و کاربر در نگارش ملزم به رعایت محدودیت‌های کمتری است. از طرفی روش پیشنهادی می‌تواند نسبت به روش‌های قبلی با دراختیارداشتن یک پایگاه داده دست‌نویس محدود، صحت مطلوب‌تری ارائه کند. با به‌کارگیری روش ارائه‌شده، دقت بازشناسی در مرحلهٔ اولیه در سطح حروف ۹۵/۹٪ و پس از بازشناسی به‌کمک مدل زبانی دقت بازشناسی به ۹۹/۳٪ ارتقا یافت. برای بهبود عملکرد الگوریتم، استفاده از الگوریتم یادگیری تقویتی برای تطبیق پذیری الگوریتم با نویسنده به‌عنوان کار آینده پیشنهاد می‌شود.

واژگان کلیدی: بازشناسی برخط، دست‌نوشته فارسی، نزدیک‌ترین همسایه، مدل زبانی، محدودیت کاربر

Online Persian Hand Writing Recognition Using Language Model and Reduction of User Writing Rules

Salman Maskanati[1] & Ahmad Keshavarz[2*]

[1]Machine Vision and Artificial Intelligence lab, Persian Gulf University, Bushehr, Iran

[2] Department of Electrical Engineering, Persian Gulf University, Bushehr, Iran

Abstract

The Joint-up, cursive form of Persian words and immense variety of its scripts, also different figures of Persian letters depending on their sitting positions in the words, have turned the Persian handwritings recognition to an intense challenge. The major obstacle of the most often recognition ways, is their inattention to sentence contexture which causes utilizing of a word with correct appearance within an incorrect sentence, when an input word is misrecognized. Sketching a solution that provides suitable analysis of sentence contexture, requires huge linguistic resources to take place as a fine representative for the chosen language to be recognized. In this article, a new method for online recognition of Persian words is presented which tries to improve recognition process by using the term contexture. In this article, the

* Corresponding author

* نویسندهٔ عهده‌دار مکاتبات

vocabularies collection of Persian language is divided into two groups. The first category is the vocabulary with all of their sub-words being supported by the database of handwritten subclasses, while these vocabulary form 68.2% of the total vocabulary, and the assumptions being scored at the recognition stage, are members of these vocabularies. The second category is the vocabulary that is not supported by the database. Obviously, if the recognition system does not support this vocabulary, it cannot recognize more than 30 percentages of the language's words. At the recognition stage, the symptoms are detected and a symptom tag is produced. Also, at this stage, using the same label, the vocabulary is also selected as the sign with the input word. (These vocabularies are chosen from those were not supported at the recognition stage). Scoring for hypotheses was done by combining recognition scores and linguistic models. The certain fact in this section is that it is impossible to calculate recognition scores due to the absence of hypothetical subheadings. Therefore, the vocabulary score being recognized in the previous steps, is used. According to the studies, it was concluded that if the word is equivalent to a member's input from a supported vocabulary, even if the result of the recognition is incorrect, in most cases the correct term is in the first four hypotheses. Usually, scores of the first few hypotheses are close to each other, and the other assumptions are far from the correct hypothesis. Since the system operates online, unnecessary computations should be avoided. Therefore, if the number of hypotheses in the recognition section are more than four hypotheses, only the first four hypotheses are calculated for the language model. To calculate the recognition score for new hypotheses, if there are fewer than four hypotheses in the recognition section, the lowest hypothesis score and otherwise the hypothesis score are considered for the recognition score of the new hypotheses. Then, as with previous assumptions, for the new hypotheses, the linguistic score is calculated, and then the final score is obtained for each hypothesis. Finally, the assumption with the highest score is considered as the system output, and the rest of the assumptions are displayed in the output to the user. Experiments show that even in the event of a mistake, the correct word is often presented as a second hypothesis in most cases, and in some cases as a third hypothesis. Also, to reduce the limits and rules that gainers compel to submit. The method demonstrated in this article includes the symptoms and morphemes framework of input handwritten are segregated and the framework of each morpheme with its symptoms is specified at first, then the symptoms of morphemes are specified and based on them a collection of words is being considered as a hypothesis. Each hypothesis is given a score by measuring the similarity to input handwritten and according to taken scores, the likely hypotheses are indicated. Then, this procedure is led to achieve hypotheses more likely by lingual models. To totalize the scores of a hypothesis, for the differences in scale of taken scores, a method of score normalization is being offered. The results demonstrate that by utilizing of a language model with an online system of handwriting recognition, a significant reduction of words recognition error rate is being achieved. In addition to error rate reduction, by taking advantages of this language model, a technique is being offered that can handle the Persian vocabulary recognition entirely. By availing the offered manner, the recognition precision at initial stage of letters level up to 95.9% and so the language model recognition up to 99.3% improved. So, using huge linguistic resources for Persian language and utilizing a language model, can improve the accuracy of recognition. For further work, reinforcement learning algorithm is suggested to adapt the algorithm for users.

Keywords: Online Recognition, Persian Handwriting, k-nearest Neighbor, Language Model, User Limitation

۱- مقدمه

امروزه با پیشرفت فناوری و گسترش روزافزون آن، ابزارهای تجاری بسیاری همچون رایانه‌های لوحی و موبایل‌های هوشمند توسعه یافته است؛ این ابزارها از ابتدای ورودشان به بازار با استقبال چشم‌گیری از سوی کاربران مواجه شدند و هرلحظه هم به محبوبیت آن‌ها بین کاربران افزوده می‌شود. یکی از حوزه‌های پژوهشی که با ظهور و توسعهٔ این ابزارها بسیار مورد توجه قرار گرفت، بازشناسی برخط دست‌نوشته بود. پژوهش‌های بین‌المللی در این زمینه به‌دلیل ساختار زبان‌های لاتین‌پایه و عدم پیچیدگی رسم‌الخط این زبان‌ها به‌سرعت به نتیجه رسید و پس از آن روند پژوهش‌ها در این زمینه بسیار کند شد؛ اما به‌دلیل تفاوت‌های ساختاری زبان‌هایی مانند فارسی که در آن کلمات به‌صورت پیوسته نوشته می‌شوند؛ پژوهش‌گران این زبان‌ها مسئولیت سنگین‌تری به دوش خواهند داشت.

در بازشناسی برخط، مختصات نقاط مسیر حرکت قلم، تعداد حرکات قلم و فشار قلم در دسترس هستند. بازشناسی برخط نوشتار به‌دلیل راحت‌تربودن نوشتن از تایپ‌کردن، عدم امکان تایپ در بعضی از موقعیت‌ها، عدم وجود یک صفحه‌کلید کامل روی رایانه‌های کوچک و سخت‌بودن تایپ حروف در بعضی زبان‌ها به‌دلیل تعداد زیاد حروف آن‌ها، مورد توجه خاصی قرار گرفته است [15].

پژوهش‌های زیادی در این زمینه برای بازشناسی دست‌نوشته‌هایی به زبان‌های مختلف از جمله لاتین-پایه، چینی، ژاپنی و عربی انجام شده است [32]، [31]، [29]، [28]، [24]، [23] و [22]. زبان فارسی نیز از این قاعده مستثنا نبوده و پژوهش‌های انجام‌گرفته در این حوزه به دو دسته کلی تقسیم می‌شوند که عبارت‌اند از بازشناسی حروف مجزا و بازشناسی کلمات پیوسته[21]، [20]، [19]، [6] و [1]. در [9] برای بازشناسی حروف مجزای فارسی از

شبکه‌های عصبی استفاده شده که در آن میزان بازشناسی درست برای ۴۱۴۴ حرف ۹۳/۹٪ گزارش شده است. در پژوهش دیگری مدل مخفی مارکوف برای بازشناسی حروف مجزای فارسی به کار گرفته شد که در آن با ارائه یک روش مبتنی بر گروه‌بندی ۲۵،۶۳ درصد خطای بازشناسی کاهش یافت [13]. در [17] و [8] برای بازشناسی برخط کلمات فارسی از روش قطعه‌بندی استفاده شده است. البته فرآیند قطعه‌بندی بسیار مستعد خطاست و می‌تواند دقت نهایی سامانه را بسیار پایین آورد [17]. لذا پژوهشگران این حوزه به دنبال نوآوری‌هایی بودند که کارایی سامانه‌های بازشناسی برخط کلمات را بهبود بخشد. برای مثال در [11] روشی برای بازشناسی زیرکلمات و در [10] برای بازشناسی کلمات فارسی ارائه‌شده است. نوآوری این روش علاوه بر جدید بودن، در اهمیت ویژه‌ای است که به بازشناسی علائم کلمات می‌دهند.

در [2] روشی برای تشخیص دست‌نوشته برخط فارسی بر مبنای شناسایی حروف سازنده زیرکلمات ارائه‌شده است که در آن الگوریتمی با روش برنامه‌نویسی پویا پیشنهاد می‌شود و درصد تشخیص حروف جدای فارسی با استفاده از ساده‌سازی و الگوریتم پیشنهادی محاسبه فاصله اصلاح برابر ۹۵/۲ ٪ شده است و درصد بازشناسی برای تشخیص زیرکلمات فارسی برای سه گزینه نخست پیشنهادی برابر ۶۳/۲۹ ٪ به دست آمد. هرچند درصد تشخیص به‌دست‌آمده نسبت به روش‌های کلی‌نگر پایین‌تر است، ولی مزیت روش ارائه‌شده در شناسایی حروف سازنده زیرکلمه نسبت به روش‌های کلی‌نگر در آن است که برای بازشناسی زیرکلمات جدید تنها کافی است که متن این زیرکلمات به فرهنگ لغت سامانه اضافه شوند، درحالی‌که در روش‌های کلی‌نگر نیاز به نمونه‌های جدید از دست‌نوشته است. در [5] یک سامانه عصبی - فازی با قابلیت آموزش هم‌زمان برای بازشناسی بر خطر زیر - کلمات فارسی ارائه شده است. روش ارائه‌شده در مقاله مذکور مبتنی بر آموزش کاربر است و در ابتدا دانشی در مورد زیرکلمه‌ها ندارد؛ لذا لازم است کاربر زیرکلمه‌های موجود در دادگان را که شامل نمونه‌های مختلف نوشتار برای هر زیر کلمه است؛ به سامانه آموزش دهد. میزان بازشناسی درست ۹۹/۸۶ درصد گزارش شده است.

در [3] روشی جدید برای بازشناسی برخط حروف مجزای فارسی ارائه شده است که با استخراج چند ویژگی ساده از دنبالهٔ نمونه‌برداری‌شده از حروف و استفاده از دسته‌بندی‌کنندهٔ ماشین بردار پشتیبان(SVM[1]) نتایج قابل قبولی را ارائه می‌دهد. نتایج تجربی این کار پژوهشی که بر اساس مجموعه داده Online-TMU صورت گرفته است، متوسط نرخ بازشناسی بدنه اصلی را ۹۴٪ نشان می‌دهد و با درنظرگرفتن پس‌پردازش‌ها بر اساس ریزحرکت‌ها این نرخ به حدود ۹۸٪ می‌رسد. در [4] روش جدید برای بازشناسی برخط زیرکلمات فارسی بر اساس کدهای زنجیره‌ای فازی و مدل تطبیق رشته با استفاده از فهرست‌های پیوندی دوطرفه ارائه شده است که علاوه‌بر کاهش پیچیدگی زمانی باعث کاهش حافظهٔ مصرفی و افزایش دقت در شناسایی زیرکلمات شده است. نرخ بازشناسی بر روی مجموعه داده استاندارد شامل یازده نمونه از هر زیرکلمه با یک فرهنگ هزار زیرکلمه‌ای ۹۱/۶۴٪ و نرخ بازشناسی کلی سامانه ۸۸/۶۷٪ است. در این مقاله برای کاهش محدودیت‌ها و قواعدی که کاربران در هنگام نوشتن مجبور به رعایت آن‌ها هستند، راه‌کارهایی مطرح خواهیم کرد که به‌سادگی قابل پیاده‌سازی هستند. در فرآیند بازشناسی، کلماتی به‌عنوان فرضیه در نظر گرفته می‌شود و به آن‌ها بر اساس میزان شباهت به ورودی امتیازی تعلق می‌گیرد؛ سپس مدل زبانی فرآیند بازشناسی کلمات دست‌نویس برخط را هدایت می‌کند و به‌این‌ترتیب سعی در کاهش خطای بازشناسی دارد.

۲- بازشناسی برخط کلمات دست‌نویس

در الگوریتم پیشنهادی این‌گونه عمل می‌شود که وقتی کاربر یک کلمه را وارد کرد، سامانه ابتدا علائم زیرکلمه‌های موجود در کلمه و مکان آن‌ها را تحلیل خواهد کرد؛ سپس سامانه باید در لغت‌نامهٔ خود به‌دنبال کلماتی بگردد که از نظر تعداد زیرکلمات و علامت‌های هر زیرکلمه مشابه الگوی تشخیص داده‌شده باشد. به‌این‌ترتیب، فضای فرضیه‌ها کاهش می‌یابد. درنهایت با استفاده از مدل زبانی بهترین گزینه از میان فرضیه‌ها انتخاب می‌شود. شکل (۱) شمای کلی روش پیشنهادی را نشان می‌دهد.

۲-۱- دریافت اطلاعات ورودی

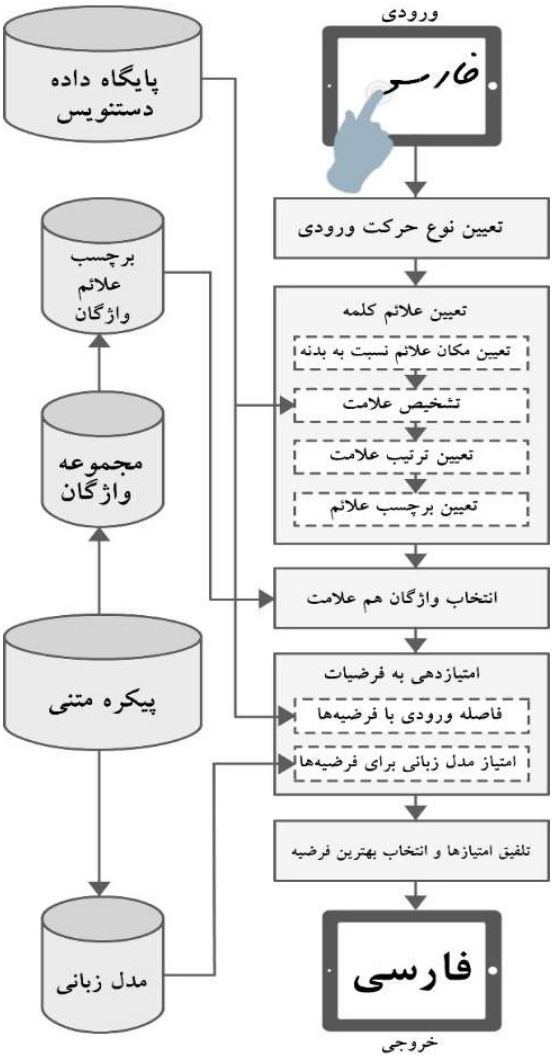
به‌طورمعمول در سامانه‌های تشخیص برخط، کاربران برای نوشتن کلمات با محدودیت‌هایی روبه‌رو هستند. در این پژوهش راه‌کارهای جدیدی برای برطرف‌کردن بخشی از این محدودیت‌ها ارائه و تا حد امکان از این محدودیت‌ها کاسته شده است. بااین‌حال هنوز هم کاربر مستلزم رعایت قواعدی

[1] Support Vector Machine

است. البته این قواعد بسیار ساده هستند و با رسم‌الخط عموم مردم سازگار است. قواعدی که کاربر برای نوشتن کلمه ورودی باید رعایت کند عبارت‌اند از:

قاعده ۱: برای نوشتن یک زیرکلمه ابتدا بدنه اصلی آن را با یک‌بار گذاشت و برداشت قلم بنویسد.

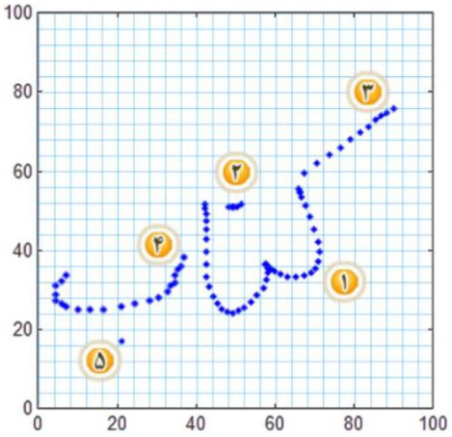
دریافت ورودی به‌گونه‌ای است که با هر حرکت قلم، مختصات نقاط مسیر حرکت قلم دریافت شود. لذا پس از نوشتن یک کلمه، تعداد حرکات قلم، تعداد نقاط نمونه‌برداری شده در هر حرکت و مختصات آن نقاط، به‌عنوان ورودی سامانه محسوب می‌شوند. در شکل (۲) نحوه دریافت کلمه «کتاب» از کاربر نمایش داده شده است. اطلاعات دریافت‌شده حاصل از نوشتن کلمه «کتاب» در جدول (۲) آمده است.



(شکل-۱): شمای کلی روش پیشنهادی
(Figure-1): Flowchart of Proposed algorithm

(جدول-۱): نحوه نوشتن علائم حروف
(Tabel-1): Writing of letter symbols

علامت	دونقطه	سه‌نقطه (نوع 1)	سه‌نقطه (نوع 2)	سرکش
نحوه نوشتن				



(شکل-۲): نقاط نمونه‌برداری شده به‌همراه ترتیب دریافت زیر حرکات
(Figure-2): Sampling points along with the order of receiving the following movements

۲-۲- تعیین نوع هر حرکت قلم و وابستگی آن به دیگر حرکات

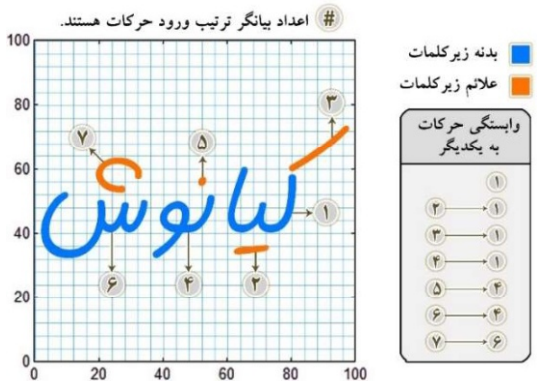
پس از واردکردن کلمه توسط کاربر، در نخستین گام باید نوع بخش‌های مختلف ورودی (حرکات قلم) مشخص شوند. طی این فرآیند هر حرکت قلم تحلیل و مشخص می‌شود که بدنهٔ یک زیرکلمه است یا علامت آن. همچنین وابستگی آن حرکت به دیگر حرکات نیز باید مشخص شود. به این معنی که اگر یکی از حرکات قلم به‌عنوان علامت شناخته شد باید مشخص شود که این علامت مربوط به کدام بدنه است. در شکل (۳) نمایی از اطلاعاتی که باید در این بخش حاصل شود آمده است.

در سایر روش‌های بازشناسی بر اساس زیرکلمه، معمولاً این مرحله فرآیند پیچیده‌ای ندارد. در مقابل یک قاعده دست و پاگیر برای کاربر تعیین می‌شود و کاربر مجبور است که برای نوشتن یک کلمه، بعد از نوشتن هر زیرکلمه علائم آن را نوشته و سپس بدنه زیرکلمه بعدی را بنویسد؛ لذا در ادامه راه‌کاری ارائه می‌شود که این محدودیت را برطرف می‌کند. در این راه‌کار باید ابتدا دامنه آن حرکت در محور طول‌ها مشخص شود. سپس با درنظرگرفتن میزان هم‌پوشانی دامنه حرکات می‌توان نوع حرکت را تشخیص داد.

(جدول-۲): اطلاعات ورودی حاصل از نوشتن کلمه «کتاب» توسط کاربر

(Tabel-2): Input information from the writing of the "کتاب" by user

حرکات قلم	تعداد نقاط	مختصات نقاط نمونه‌برداری
حرکت اول	57	$\{\binom{66}{58}\binom{67}{57}\dots\binom{42}{53}\}$
حرکت دوم	9	$\{\binom{47}{53}\binom{48}{53}\dots\binom{55}{54}\}$
حرکت سوم	11	$\{\binom{97}{76}\binom{89}{84}\dots\binom{66}{58}\}$
حرکت چهارم	21	$\{\binom{37}{39}\binom{36}{37}\dots\binom{7}{36}\}$
حرکت پنجم	1	$\{\binom{21}{17}\}$



(شکل-۳): تعیین نوع هر حرکت قلم و وابستگی آن به دیگر حرکات

(Figure-3): Determine the type of pen movement and its dependence on other movements

(جدول-۳): وضعیت‌های مختلف هم‌پوشانی حرکت جاری با بدنه‌ها

(Table-3): Different situations of current movement overlapping with framework

ردیف	وضعیت‌های مختلف هم‌پوشانی	تصویر وضعیت‌ها (آبی = بدنه، قرمز = حرکت جاری)
۱	سطح بدنه و حرکت جاری هم‌پوشانی نداشته باشند.	
۲	حرکت جاری کوچک‌تر از بدنه و کاملاً با آن هم‌پوشانی داشته باشد.	
۳	حرکت جاری با بدنه مقداری هم‌پوشانی داشته باشد.	
۴	حرکت جاری با چندین بدنه هم‌پوشانی داشته باشد.	
۵	حرکت جاری بزرگ‌تر از بدنه و کاملاً با آن هم‌پوشانی داشته باشد.	

حرکت نخست قلم همیشه به‌عنوان بدنه زیرکلمهٔ نخست در نظر گرفته می‌شود و به حرکت دیگری وابستگی ندارد. برای سایر حرکات باید میزان هم‌پوشانی آن‌ها با بدنه‌هایی که تاکنون تشخیص داده شده است بررسی شود. دامنه حرکات ممکن است در چند موقعیت هم‌پوشانی مختلف با بدنه‌ها قرار گیرند. این وضعیت‌ها در جدول (۳) آمده است. در تصاویر این جدول رنگ آبی نمادبدنه و رنگ قرمز نماد حرکت جاری است.

۱) اگر در مقایسه حرکت جاری با تمامی بدنه‌ها، با هیچ بدنه‌ای هم‌پوشانی وجود نداشت، حرکت جاری به‌عنوان بدنه تشخیص داده خواهد شد و به آخرین بدنهٔ ورودی قبل از خود، وابسته خواهد بود.

۲) اگر حرکت جاری به‌طورکامل تحت پوشش یک بدنه باشد و با دیگر بدنه‌ها هیچ هم‌پوشانی‌ای نداشته باشد، به‌عنوان علامت آن بدنه محسوب و وابسته به آن خواهد شد.

۳) اگر حرکت جاری فقط با یکی از بدنه‌ها مقداری هم‌پوشانی داشته باشد، چند حالت مختلف اتفاق خواهد افتاد. تصمیم‌گیری در مورد این حالت‌ها به این‌گونه خواهد بود که اگر بدنه‌ای که با حرکت جاری هم‌پوشانی دارد، آخرین بدنهٔ واردشده و حرکت جاری در سمت چپ بدنه باشد و میزان هم‌پوشانی کمتر از ۷۰ ٪ باشد، حرکت جاری به‌عنوان بدنه در نظر گرفته می‌شود و به آخرین بدنهٔ ورودی وابسته خواهد بود. در بقیهٔ حالات، حرکت جاری به‌عنوان علامت محسوب شده و به بدنه‌ای که با آن هم‌پوشانی دارد، وابسته است. مقدار ۷۰٪ با آزمون و خطا به‌دست‌آمده است.

۴) در هنگامی‌که حرکت جاری با چندین بدنه هم‌پوشانی دارد، به‌عنوان علامت در نظر گرفته شده ولی اینکه به کدام بدنه وابسته است، باید بررسی شود؛ لذا ابتدا بر روی حرکت جاری یک بازشناسی اولیه انجام می‌شود که طی آن مشخص می‌شود حرکت واردشده «دونقطه»، «سه‌نقطه»، «سرکش»، «مد» یا «همزه» است. نحوهٔ فرآیند بازشناسی علائم در ادامه به‌تفصیل شرح داده خواهد شد. پس از بازشناسی اگر حرکت جاری، «همزه» یا «مد» تشخیص داده شد، حرکت جاری به بدنه «الف» وابسته می‌شود. اکنون اگر حرکت جاری «دونقطه» یا «سه‌نقطه» بازشناسی شده باشد، به بدنه‌ای نسبت داده می‌شود که بیشترین هم‌پوشانی را با آن بدنه دارد؛ و درنهایت اگر حرکت جاری «سرکش» تشخیص داده شد،

از میان بدنه‌هایی که با آن‌ها هم‌پوشانی دارد، به سمت چپ‌ترین بدنه وابسته خواهد بود.

۵) در این حالت دامنهٔ حرکت از بدنه بزرگ‌تر است. این حالت فقط برای «مد» و «همزه» اتفاق می‌افتد؛ لذا حرکت جاری به‌عنوان علامت محسوب شده و وابسته به تک‌بدنه‌ای است که با آن هم‌پوشانی دارد.

۲-۳- تعیین مکان علائم نسبت به بدنه

در نخستین گام از بازشناسی علامت، باید مشخص شود که علامت در بالای بدنه یا پایین آن قرار گرفته است. برای این منظور ابتدا چارچوب علامت و بدنه‌ای که علامت به آن وابسته است تعیین و مختصات مرکز هر چارچوب مشخص می‌شود. اگر مرکز چارچوب علامت بالاتر یا پایین‌تر از چارچوب بدنه باشد، موقعیت آن به‌ترتیب بالا یا پایین در نظر گرفته می‌شود. در غیر این صورت روی نقاط بدنهٔ مربوطه حرکت می‌کنیم. برای هر جفت نقطهٔ متوالی در مسیر حرکت اگر روی محور طول‌ها، مرکز چارچوب علامت بین یا سمت راست این جفت نقطه قرار گیرد، درصورتی‌که مرکز چارچوب علامت بالاتر از خط واصل این دونقطه باشد، موقعیت علامت جاری بالا در نظر گرفته می‌شود و درصورتی‌که مرکز چارچوب علامت پایین‌تر از خط واصل این دونقطه باشد، موقعیت پایین برای آن در نظر گرفته می‌شود [11].

۲-۴- پیش‌پردازش

طی این فرآیند جزییات غیرمفید از داده‌های خام حذف می‌شوند و تمامی داده‌ها در یک چارچوب مشابه برای انجام پردازش‌های بعدی آماده می‌شوند.

۲-۴-۱- یکسان‌سازی تعداد و فاصله بین نقاط

یک سامانهٔ کارا باید مستقل از نویسنده باشد؛ لذا برای حذف ویژگی سرعت نوشتن و همچنین حذف لغزش‌های قلم، لازم است فاصله بین نقاط و همچنین تعداد نقاط نمونه‌برداری‌شده برابر باشد. برای این منظور ابتدا یک درون‌یابی انجام می‌شود و یک منحنی به‌دست می‌آید که تخمینی از شکل ورودی است؛ سپس بر روی این منحنی N نقطه با فاصلهٔ مساوی اختیار می‌شود که تعداد این نقاط به نوع پردازش بستگی دارد. در این پژوهش برای علائم ده نقطه و برای بدنه‌ها پنجاه نقطه انتخاب شده است [11].

۲-۴-۲- یکسان‌سازی ابعاد

ابعاد نوشتن نیز یکی دیگر از عواملی است که به سلیقهٔ کاربران بستگی دارد؛ لذا برای اینکه این عامل در بازشناسی دخیل نباشد باید همهٔ ورودی‌های دریافت‌شده، طی فرآیندی در یک مقیاس خاص قرار گیرند. در این پروژه از این پیش‌پردازش فقط برای بدنهٔ زیرکلمات استفاده می‌شود که طی آن تمامی بدنه‌ها در ابعاد ۵۰×۵۰ قرار می‌گیرند.

۲-۵- استخراج ویژگی

پس‌ازاینکه داده‌های خام ورودی پیش‌پردازش شد، فرآیند استخراج ویژگی را بر روی نقاط پیش‌پردازش شده می‌توان اعمال کرد. در این مقاله از دو ویژگی بهره برده‌ایم، یکی ویژگی زاویهٔ بین نقاط متوالی و دیگری ویژگی مختصات نقاط که در ادامه شرح داده خواهند شد.

۲-۵-۱- ویژگی زاویهٔ بین نقاط

محاسبهٔ ویژگی زاویهٔ بین نقاط به این‌گونه است که نقاط متوالی را دوبه‌دو در نظر گرفته و زاویه خط واصل جفت‌نقطه‌های مجاور با محور طول‌ها محاسبه می‌شود. برای این منظور ابتدا شیب خط واصل جفت‌نقطه‌های مجاور از رابطه (۱) محاسبه می‌شود.

۲-۵-۲- ویژگی مختصات نقاط

درصورتی‌که در بخش پیش‌پردازش ابعاد ورودی یکسان‌سازی شود، مختصات نقاط نیز می‌تواند به‌عنوان یک ویژگی، مفید باشند. ویژگی مختصات نقاط همان مقادیر محورهای X و Y است که بدون تغییر در یک بردار $N \times 2$ درایه‌ای ذخیره می‌شوند.

$$m = \frac{\Delta y}{\Delta x} = \frac{(y_A - y_B)}{(x_A - x_B)} = \tan\theta \qquad (۱)$$

اکنون زاویه خط واصل A و B با استفاده از شیب خط و از رابطهٔ زیر به‌دست می‌آید.

$$\theta = \tan^{-1} m \qquad (۲)$$

۲-۶- تعیین علامت

برای تعیین علامت ورودی، چارچوب آن در محور مختصات استخراج و اگر عرض و ارتفاع آن از یک آستانه خاص کمتر بود، علامت به‌عنوان «نقطه» شناسایی می‌شود. اگر علامت

موردبررسی به‌عنوان «نقطه» بازشناسی نشد، باید شرط «دسته» بودن علامت بررسی شود. علامت جاری در صورتی به‌عنوان «دسته» شناسایی خواهد شد که ارتفاع علامت، سه برابر بزرگ‌تر از پهنای آن علامت باشد. شروط مربوط به «نقطه» و «دسته»، چون بر اساس چهارچوب ورودی عمل می‌کند، باید قبل از هرگونه پیش‌پردازش و فقط با تحلیل داده‌های خام ورودی، انجام شود. اگر علامت ورودی به‌عنوان نقطه یا دسته شناسایی نشد، باید داده‌های مربوط به آن علامت مورد پیش‌پردازش قرار گیرد. برای پیش‌پردازش در این بخش، یکسان‌سازی تعداد و فاصله بین نقاط در نظر گرفته شده است. در این پیش‌پردازش مقدار N برابر ده در نظر گرفته شده است که از این نقاط ویژگی زاویه استخراج می‌شود. با استخراج این ویژگی یک بردار با ۹ درایه خواهیم داشت که هر مقدار بیان‌گر زاویهٔ خط واصل دو نقطهٔ مجاور و محور طول‌هاست. از هر علامت یکصد نماینده در پایگاه داده موجود می‌باشد. بردار زاویهٔ استخراج‌شده باید با بردار زاویهٔ هرکدام از نماینده‌های علائم مقایسه شود. طی این مقایسه فاصلهٔ این بردار با تمام نمایندگان علائم محاسبه می‌شود. مقایسه به‌صورت زیر انجام می‌شود. فاصلهٔ زوایا d_a از رابطهٔ

$$d_a(i.j) = \sum_{k=1}^{N-1} \min\left(\left|A_i^k - A_j^k\right|, 360 - \left|A_i^k - A_j^k\right|\right) \quad (۳)$$

به‌دست می‌آید. A_i^k عنصر k ام از بردار A_i است که بردار زاویه علامت ورودی است [12]. بردار A_j نیز بردار زاویه نمونهٔ j ام است که برای پانصد نمونه محاسبه می‌شود. درنهایت علامتی که نماینده‌ای از آن کمترین فاصله را با علامت ورودی دارد به‌عنوان نتیجه بازشناسی انتخاب می‌شود.

۱) برای تمامی سرکش‌ها مراحل ۲ تا ۹ را اجرا کن.
۲) اگر تمامی سرکش‌ها به‌صورت دوبه‌دو با یکدیگر مقایسه شده‌اند برو به مرحله ۱۰. در غیر این صورت برو به مرحله ۳.
۳) دو سرکشی که تابه‌حال با یکدیگر مقایسه نشده‌اند را در نظر بگیر.
۴) سرکش بزرگ‌تر و کوچک‌تر را بر اساس پهنای آن‌ها مشخص کن.
۵) اگر سمت راست‌ترین نقطه سرکش کوچک‌تر در محدوده پهنای سرکش بزرگ‌تر بود، مراحل ۴ تا ۵ را اجرا کن در غیر این صورت برو به مرحله ۲.
۶) اگر پهنای سرکش کوچک‌تر، کمتر از ۶۰ درصد پهنای سرکش بزرگ‌تر بود، بود مراحل ۷ تا ۹ را اجرا کن، در غیر این صورت برو به مرحله ۲.
۷) اگر پهنای سرکش کوچک‌تر بیش از ۷۰ درصد با پهنای سرکش بزرگ‌تر همپوشانی داشت برو به مرحله ۸ درغیراین‌صورت بروبه مرحله ۲.
۸) سرکش بزرگ‌تر را به‌عنوان علامت حرف «گ» در نظر بگیر.
۹) سرکش کوچک‌تر را از مجموعه علائم حذف کن.
۱۰) پایان

(شکل-۴): شبه‌کد تشخیص علائم حرف «گ»

(Figure-4): Pseudo-code of recognition of the letter "گ"

۲-۷- تشخیص علائم حرف «گ»

علامت حرف «گ» ترکیبی از دو «سرکش» است؛ لذا باید علائمی که تا این مرحله شناسایی شده‌اند، بررسی شوند و اگر در میان علائم بیش از یک «سرکش» وجود داشت احتمال این می‌رود که این دو سرکش علائم حرف «گ» باشند. برای اینکه بتوانیم این حالات را تشخیص دهیم از شبه‌کد شکل (۴) پیروی می‌شود.

۲-۸- تشخیص سه‌نقطه ترکیبی

این فرآیند برای هر زیرکلمه به‌صورت مستقل انجام می‌شود. برای تشخیص سه‌نقطه بالا، ابتدا بررسی می‌شود که در یک زیرکلمه علامت دو نقطه بالا و یک نقطه بالا وجود داشته باشد؛ سپس به‌ازای هرکدام از دونقطه‌ها، تمامی تک‌نقطه‌ها بررسی می‌شود.

محدودهٔ پهنای علامت دونقطه در نظر گرفته می‌شود. اگر مرکز تک‌نقطه در محدوده دوچهارم میانی علامت دونقطه قرار داشت، این دو علامت با یکدیگر ترکیب‌شده و علامت «سه‌نقطه بالا» را به‌وجود می‌آورد. این فرآیند برای دونقطه‌ها و تک‌نقطه‌های پایین نیز به‌صورت جداگانه بررسی می‌شود.

۲-۹- تعیین ترتیب علائم

نویسندگان ترتیب خاصی را در موقع نوشتن علائم رعایت نمی‌کنند. برای اینکه به‌ترتیب واقعی علائم برسیم باید آن‌ها را از نظر مکانی مرتب کنیم. برای این منظور از مؤلفه X سمت چپ‌ترین نقطه در ورودی‌ها استفاده می‌شود؛ یعنی X_{min} برای هر علامت محاسبه‌شده و به‌ترتیب نزولی مرتب می‌شود [11].

۲-۱۰- تولید برچسب علائم برای کلمه

در مراحل قبلی، تعداد بدنه‌ها و علائم هر بدنه مشخص شد. اکنون باید بر اساس علائم، برچسب زیرکلمات را تعیین و با ترکیب آن‌ها برچسب کلمه را تولید کنیم. برای تعیین برچسب از (جدول (۴) پیروی می‌شود.

(جدول-۴): نمایندگان علائم موجود در زیرکلمات
(Tabel-4): Representatives of existing symptoms at sub-words

علامت	نماینده	گروه علائم
یک‌نقطه پایین	ب	«ب»، «ج»
یک نقطه بالا	ن	«ن»، «خ»، «ذ»، «ز»، «ض»، «غ»، «ف»
دونقطه پایین	ی	«ی» (فقط در ابتدا و وسط زیرکلمه)
دونقطه بالا	ت	«ت»، «ق»
سه‌نقطه پایین	پ	«پ»، «چ»
سه‌نقطه بالا	ث	«ث»، «ژ»، «ش»
یک سرکش	ک	«ک»
دو سرکش	گ	«گ»
دسته	ط	«ط»، «ظ»
همزه میانی	ئ	«ئ» (علائم «ٔ» و «ئ»)، «ؤ»
همزه پایین	إ	«إ»
همزه بالا	أ	«أ»
الف	ا	«ا»
الف با کلاه	آ	«آ»
بدون علامت	م	برای زیرکلماتی که هیچ علامتی ندارند

برای تولید برچسب یک زیرکلمه در ابتدا «الف» بودن بدنه زیرکلمه بررسی می‌شود. درصورتی‌که ارتفاع بدنه سه‌برابر پهنای آن باشد، بدنه به‌عنوان «الف» در نظر گرفته می‌شود. درصورتی‌که بدنهٔ «الف» فاقد علامت باشد، برچسب «ا» را برای آن در نظر می‌گیریم؛ در غیر این صورت بسته به علامت آن برچسب «إ»، «أ» یا «آ» به آن زیرکلمه نسبت داده می‌شود. برای بدنه‌های غیر «الف» اگر فاقد علامت باشد، برچسب «م» به آن تعلق می‌گیرد؛ اما درصورتی‌که دارای علامت باشد، بسته به علامت‌های تشخیص‌داده‌شده، برای هر علامت یکی از نمایندگان علائم انتخاب خواهد شد. به‌عنوان برچسب علائم کلمه «آشپزی» بر اساس جدول مذکور به‌صورت«آ_ثپن_م» است.

۲-۱۱- انتخاب واژگانی کوچک هم‌علامت با کلمه

از میان واژگان، فقط آن واژه‌هایی انتخاب خواهند شد که برچسب علائم آن‌ها با برچسب علائم ورودی یکسان باشد. به‌این‌ترتیب فضای فرضیات بسیار محدود می‌شود. به‌عنوان‌مثال برای کلمهٔ «کتاب» که برچسب آن «کت_ب» است، فضای فرضیات از ۶۰۴۰۵ واژه به چهار واژه کاهش می‌یابد.

۲-۱۲- بازشناسی کلمه ورودی

با مشخص‌شدن فضای فرضیه‌ها باید به هر فرضیه بر اساس میزان شباهت با ورودی امتیازی تعلق گیرد. اگر در فضای فرضیات فقط یک فرضیه موجود باشد، بازشناسی خاتمه می‌یابد؛ اما گر این فضا بیش از یک فرضیه داشته باشد باید بدنهٔ هر زیرکلمه از ورودی با بدنهٔ متناظر آن در فرضیه مقایسه شود.

۲-۱۲-۱- محاسبه فاصلهٔ بدنه ورودی با بدنهٔ فرضیه‌ها

در این قسمت فاصلهٔ هر بدنه از ورودی با بدنهٔ هم‌علامت با خودش در فضای فرضیات مقایسه می‌شود. برای این منظور ابتدا باید بدنه ورودی پیش‌پردازش شود، پیش‌پردازش‌های یکسان‌سازی تعداد و فاصله بین نقاط و یکسان‌سازی ابعاد بر روی بدنه اعمال و سپس استخراج ویژگی انجام می‌شود. ویژگی‌های درنظرگرفته‌شده برای این بخش، ویژگی زاویه و ویژگی مختصات نقاط است. اکنون باید ویژگی‌های استخراجی با ویژگی‌های بدنهٔ زیرکلمه‌های فرضیات که در پایگاه داده موجود است، مقایسه شود. فاصله میان این ویژگی‌ها با روابط زیر محاسبه خواهد شد. فاصلهٔ ویژگی مختصات نقاط d_c از رابطه

$$d_c(i.j) = \sum_{k=1}^{N} \sqrt{\left(x_i^k - x_j^k\right)^2 + \left(y_i^k - y_j^k\right)^2} \qquad (۴)$$

به‌دست می‌آید که در آن x_i^k مؤلفهٔ x نقطهٔ k ام از بدنهٔ هنجارسازی‌شدهٔ ورودی و $N = 50$ است. x_j^k نیز مؤلفهٔ x نقطهٔ k ام از بدنهٔ فرضیه j ام است (رضوی؛ کبیر ۱۳۸۴). فاصلهٔ زوایا d_a نیز از رابطه (۳) محاسبه می‌شود و در آن A_j بردار زاویه فرضیهٔ j ام است.

۲-۱۳- امتیازدهی به فرضیه‌ها

فرض می‌شود که H فرضیهٔ متفاوت برای تصمیم‌گیری وجود دارد. برای هرکدام دو مقدار فاصله به‌ازای زوایا و مختصات به‌دست آمده است. در این بخش به هر فرضیه یک امتیاز تعلق می‌گیرد، این امتیاز با ترکیب ویژگی‌های زوایا و مختصات تولید می‌شود. امتیاز نهایی از رابطهٔ زیر محاسبه می‌شود:

$$d(i.j) = d_c(i,j) + \alpha d_a(i,j) \qquad (۵)$$

که در آن d_c فاصله‌ی نقاط و d_a فاصله‌ی زوایای بین نقاط هستند.

۲ -۱۴- توسعهٔ کلمات تحت پوشش پایگاه داده

در بخش بازشناسی برای مقایسه کلمات، باید بدنه زیرکلمات آن‌ها باهم مقایسه شود. تعداد واژگان برابر ۸۸۵۶۳ است که ۷۰۹۲۰ واژه از پیکره متنی دکتر بی‌جن‌خان و بقیه از پایگاه اینترنتی دبیرخانه شورای عالی اطلاع‌رسانی استخراج شده‌اند. این مجموعه دارای ۱۲۶۰۹ زیرکلمهٔ مختلف است. در این مقاله از پایگاه‌داده زیرکلمات دست‌نویس دانشگاه تربیت مدرس [12] استفاده شده است که فقط هزار زیرکلمه پرکاربرد را در بر دارد، به‌طورطبیعی فقط مجاز به استفاده از آن واژگانی هستیم که تمامی زیرکلمات تشکیل‌دهنده آن کلمه، در پایگاه داده زیرکلمات دست‌نویس موجود باشد. واژگانی که زیرکلمات آن‌ها توسط پایگاه دادهٔ دست‌نویس پشتیبانی می‌شوند، برابر است با ۴۸۴۲۲ کلمه که معادل ۵۴/۷ ٪ از کل واژگان است. ازآنجایی‌که در این پژوهش و برخی پژوهش‌های دیگر در این زمینه[10]،[9]، فقط از بدنهٔ زیرکلمات برای بازشناسی استفاده می‌شود، لذا در مجموعه واژگان زیرکلمه‌هایی است که به‌صورت مستقیم در پایگاه داده وجود ندارند؛ ولی بدنهٔ آن به‌واسطهٔ زیرکلمات دیگر تحت پوشش پایگاه داده قرار می‌گیرد. به‌عنوان‌مثال کلمه «آبشار» به دلیل اینکه زیرکلمه «بشا» در پایگاه داده وجود ندارد، پشتیبانی نمی‌شود، حال‌آنکه زیرکلمهٔ «نشا» توسط پایگاه داده پشتیبانی می‌شود؛ لذا می‌توان از بدنهٔ زیرکلمه «نشا» در هنگام بازشناسی کلمه «آبشار» استفاده کرد.

۳- پیکرهٔ متنی زبان فارسی

طبق تعریف، پیکره حجم زیادی از داده‌های زبانی است که بر اساس معیارهای مشخص برای هدف معینی جمع‌آوری و ذخیره شده به‌طوری‌که نماینده زبان یا گویش مورد مطالعه باشد [30].

در این پژوهش برای تولید مدل زبانی از پیکره بی‌جن‌خان [26] که یک پیکرهٔ متنی استاندارد است، استفاده شده است. این پیکره تقریباً شامل ۲,۶ میلیون (۲،۵۹۷،۹۳۷) کلمه برچسب خورده است. مجموعه برچسب به‌کاررفته در این پیکره متنی، در ابتدا ۵۵۰ برچسب و سپس به ۴۰ برچسب مختلف کاهش یافت که هر یک از آن‌ها بیان‌کننده یک گونه صرفی است. این پیکره شامل متون رسمی، غیررسمی و محاوره‌ای است.

۳-۱- یک‌دست‌سازی املایی پیکرهٔ متنی

یک‌دست‌سازی املایی ازآن‌جهت لازم است که بعضی کلمات با دو یا چند املای مختلف در پیکره متنی نوشته شده‌اند. به‌عنوان‌مثال پیشوند «می» و پسوند «ها» در ابتدا و انتهای کلمات، به سه صورت (جدول ۵) ممکن است در متن فارسی دیده شود یا کلمات «مسئول»، «مجموعهٔ» و «پاییز» به‌صورت‌های مختلفی در پیکرهٔ متنی دیده می‌شود [7] (جدول۶).

(جدول-۵): چندامل‌ایی‌نوشتن پسوندها و پیشوندها در زبان فارسی

(Tabel-5): Multi spelling for writing extensions and prefixes in Persian language

جدا بافاصله	جدا بدون فاصله	متصل
کتاب ها	کتاب‌ها	کتابها
می روند	می‌روند	میروند

(جدول-۶): چندامل‌ایی نوشتن کلمات «مسئول»، «مجموعه» و «پاییز»

(Tabel-6): Multi spelling of «مسئول»، «مجموعه» و «پاییز»

مسوول	مسؤول	مسئول
مجموعه	مجموعهٔ	مجموعه ی
پاییز	پائیز	

بنابراین نویسه‌های به‌کاررفته در دو کلمهٔ یکسان، ممکن است، متفاوت باشند. این باعث می‌شود که هنگام شمردن کلمات، دو کلمهٔ یکسان ولی با املای متفاوت به‌عنوان دو کلمه مختلف در نظر گرفته شوند [7]. در ادامه موارد مختلفی از یک‌دست‌سازی املایی آمده که به‌صورت دستی یا با استفاده از نرم‌افزار «واژه‌آرا» انجام‌گرفته است.

- تبدیل نویسه‌های «ی» به «ی» و «ک» به «ک».
- حذف «ی» اضافه از کلماتی مانند «کلمه ی».
- حذف فتحه (نویسه ـَ) و کسره (نویسه ـِ) و ضمه (نویسه ـُ).
- حذف تنوین‌های از نوع «ـً»، در مواردی نیز به‌صورت «ا"» دیده‌شده‌اند که آن‌ها را حذف کردیم.
- حذف نویسه تشدید «ـّ».

- چسباندن پیشوندها بـه ابتـدای کلمـه.ماننـد «مـی»، «بی»، «نمی» .
- چسباندن پیشوند «هم» در کلماتی مانند «هم‌چنین» به ابتدای کلمات.

۳-۲- تصحیح غلط‌های املایی

در بررسی پایگاه داده تعداد زیادی غلط‌های املایی مشـاهده شد که اصـلاح گردیـد. ظـاهراً در هنگـام جمـع‌آوری پیکـره متنی، برخی کلمات دست‌خوش تغییر شده‌اند، بـه‌عنوان‌مثال در فایلی کـه تارنمـای مرجـع پیکـره متنـی ارائـه کـرده در بسیاری از کلمات، بـه جـای «ه»، نویسـه «ث» قـرار گرفتـه است. مانند «فاصلث»، «وسیلث» و «مرحلث». همچنـین در بسیاری موارد، دو نویسه «اا» جایگزین «ائ» و «ای» شده مانند «رااه»، «ملاکه» و «عقااد». چند نمونه از غلط‌هـای املایـی پیکره متنی در جدول (۷) آمده است.

(جدول-۷): نمونه‌هایی از غلط‌های املایی پیکره متنی
(Tabel-7): Examples of spelling mistakes in text

واژه اصلی	واژه اصلاح‌شده	واژه اصلی	واژه اصلاح‌شده
توطاه	توطئه	راوس	رأس
پاایز	پاییز	شاونات	شئونات

لازم به ذکر اسـت کـه تمـامی اصـلاحات بـر اسـاس دستور خط و زبان فارسی مصـوب فرهنگسـتان زبـان و ادب فارسی صورت گرفته است[14]،[18].

۴- استخراج اطلاعات آمـاری از پیکـره متنی

آمارهای مورد نیاز برای طراحی مدل زبانی به شرح زیر است [7].

۱) تعداد رخداد هرکدام از کلمات؛
۲) تعداد رخداد هر دوتایی از کلمات پیکره‌ی متنی؛

این آمار فقط برای دوازده‌هزار کلمـه پرکـاربرد به‌دسـت می‌آید. سـایر کلمـات به‌صـورت یـک کلمـه «خـارج از واژگان» در نظر گرفته می‌شوند.

۳) تعداد رخداد هر سه‌تایی از کلمات موجود در پیکره متنی؛

مانند آمار تعداد رخداد دوتـایی کلمـات، ایـن آمـار نیـز فقط برای دوازده‌هزار کلمـه پرکـاربرد به‌دسـت می‌آیـد. ازآنجاکه بسیاری از ترکیب‌های سه‌تایی از ایـن کلمـات در پیکره متنی موجود نیسـت، فقـط آمـار ترکیب‌هـای سه‌تایی موجود به‌صورت یک فهرست ذخیره می‌شود.

۴) تعداد رخداد هر برچسب در پیکره متنی؛
۵) تعداد رخداد هر دوتایی از برچسب‌های موجـود در پیکره متنی؛

این آمار به‌صورت یک ماتریس۴۰×۴۰ ذخیره می‌شود.

۶) تعداد رخـداد هـر سـه‌تایی از برچسـب‌های پیکـره متنی؛

این آمـار به‌صـورت یـک مـاتریس ۴۰×۴۰×۴۰ ذخیـره می‌شود.

۷) تعداد رخداد هر برچسب به‌ازای هر کلمه.

هـر کلمـه می‌توانـد بنـا بـه نقشـی کـه در جملـه دارد، برچسـب‌های متفـاوتی بگیـرد؛ لـذا آمـار برچسـب‌های مختلف برای هر کلمـه از پیکـره متنـی اسـتخراج و در ماتریس با ابعاد ۱۲۰۰۱×۴۰ ذخیره می‌شـوند. بـه‌عنوان مثال برای کلمه «زیبا» که می‌تواند اسم و صفت باشـد، این آمار به‌صورت جدول (۸) استخراج می‌شود [7].

(جدول-۸): انواع مختلف برچسب برای کلمه زیبا و تعداد آن‌ها
(Tabel-8): Different types of labels for the "زیبا" word and their number

انواع برچسب برای کلمه زیبا	تعداد
N_SIGN_PR	17
ADJ_SIM	420

۵- مدل‌های زبانی $n-gram$

ساده‌ترین و پرکاربردترین مدل زبانی آمـاری، مـدل‌های $n-gram$ هستند. به‌صورت کلی این مدل‌ها بر اسـاس احتمـال رخـداد یـک کلمـه پـس از دنبالـه‌ای از $n-1$ کلمـه عمـل می‌کند. احتمال دنباله لغات در حالت کلـی برابـر اسـت بـا $W=w_1,w_2,w_3,\dots,w_m$این احتمال با رابطـهٔ زیـر تعریـف می‌شود:

$$P(W)=P(w_1w_2\dots w_m)=\prod_{i=1}^{m}P(w_i|w_1\dots w_{i-1}) \quad (۶)$$

درعمل تعداد کلمات قبلی به $n-1$ کلمه محدود می‌شـود و مدل حاصـل $n-gram$ نامیـده می‌شـود. وقتـی m بـزرگ باشد، محاسبه احتمال بالا بسیار مشکل و درعمل غیـرممکن است. مقادیر n بـه‌طورمعمول بـین یـک تـا پـنج اسـت و بـا نام‌های زیر شناخته می‌شوند.

- $n = 1$: (unigram) monogram
- $n = 2$: bigram
- $n = 3$: trigram
- $n = 4$: (4-gram)quadrigram
- $n = 5$: $5 - \text{gram}$

برای $n = 1$ (مدل monogram):

$$P(W) = P(w_1)P(w_2)P(w_3) \dots P(w_m) \quad (۷)$$

برای $n = 3$ (مدل trigram):

$$P(W) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2) \dots P(w_m|w_{m-2}w_{m-1}) \quad (۸)$$

مدل‌های $n - gram$ با استفاده از شمارش دنبالهٔ کلمات در یک پیکره متنی بزرگ به‌دست می‌آیند. ابتدا تمام انواع کلمات پیکره شمارش و یک مجموعه واژگان شامل V کلمه از کلمات پرکاربرد (و سایر کلمات مورد نظر) تعیین می‌شود. سایر کلمات همگی با یک نماد مشخص به‌عنوان کلمه خارج از واژگان (OOV) جایگزین می‌شوند؛ سپس پیکره از ابتدا تا انتها پیمایش شده و تمام ترکیبات دوتایی، سه‌تایی، ... و nتایی از کلمات واژگان (و همچنین نماد OOV) شمارش می‌شود.

تعداد پارامترهای مدل $n - gram$ با افزایش n به‌طور نمایی رشد می‌کند.

تعداد پارامترهای مدل bigram: V^2
تعداد پارامترهای مدلtrigram: V^3
تعداد پارامترهای مدلn-gram: V^n

به‌طورمعمول V (تعداد واژگان) از مرتبهٔ چند هزار است بنابراین تعداد پارامترهای مدل $n - gram$ بسیار زیاد است. در زبان‌هایی مانند فارسی به‌دلیل کمبود حجم داده آموزشی، n به ۱، ۲ و یا به ۳ محدود می‌شود؛ ولی در زبان‌هایی که حجم داده‌ها بسیار زیاد است $n - gram$ های بزرگ‌تری هم تولید شده است. به‌عنوان‌مثال شرکت «گوگل» برای زبان انگلیسی موفق به تولید $5 - gram$ شده است. بسته به این‌که احتمال‌های مذکور از بین کلمات یا از بین کلاس‌هایی از کلمات استخراج شوند، مدل‌های $n - gram$ متفاوتی وجود خواهد داشت.

در مورد مدل monogram، مدل بر اساس خود کلمات استخراج شده است؛ ولی در مورد مدل‌های bigram و trigram، می‌توان مدل را هم از خود کلمات استخراج و هم می‌توان کلمات را دسته‌بندی کرده و مدل را بر اساس دسته‌های کلمات استخراج کرد [7]. در این نوشتار دسته‌بندی‌ها بر اساس پاره‌گفتار صورت گرفته است.

۵-۱- مدل *n-gram* مبتنی بر کلمه

ساده‌ترین مدل زبانی $n - gram$ مدل $n - gram$ مبتنی بر کلمه است. در این مدل، احتمال شرطی رخداد هر کلمه پس از هر رشته $n - 1$ کلمه‌ای مورد توجه قرار می‌گیرد. مدل‌های monogram، bigram و trigram مبتنی بر کلمه به‌ترتیب با روابط زیر به دست می‌آیند[7].

$$P(w) = \frac{N(w)}{N_{total}} \quad (۹)$$

$$P(w_2|w_1) = \frac{N(w_1w_2)}{N(w_1)} \quad (۱۰)$$

$$P(w_3|w_1w_2) = \frac{N(w_1w_2w_3)}{N(w_1w_2)} \quad (۱۱)$$

در روابط بالا $N(w)$، $N(w_1w_2)$، $N(w_1w_2w_3)$ به ترتیب آمارهای monogram، bigram و trigram کلمات و N_{total} تعداد کل کلمات پیکره متنی است.

درعمل بسیاری از این پارامترها صفر هستند؛ یعنی دنباله کلمات مربوط به آن‌ها در پیکره متنی رخ نداده است. این اتفاق به دو دلیل رخ می‌دهد: یکی به‌دلیل کم‌بودن حجم پیکره متنی و دیگری اینکه بعضی از دنباله کلمات در زبان مجاز نیستند. بنابراین نحوهٔ ذخیرهٔ مدل‌های $n - gram$ در حافظهٔ رایانه به‌صورت ماتریس sparse است.

۵-۲- مدل $n - gram$ مبتنی بر پاره‌گفتار

برای استخراج آمار bigram و trigram معتبر نیاز به حجم عظیمی از داده‌های متنی است. پیکره متنی هرچقدر هم که بزرگ باشد؛ باز هم بسیاری از ترکیبات دوتایی و سه‌تایی کلمات را پوشش نمی‌دهد؛ به همین دلیل بسیاری از آمارهای bigram و trigram صفر و یا خیلی کوچک هستند. برای برطرف کردن نسبی این مشکل و به‌دست آوردن آمارهای معتبرتر، به‌طورمعمول کلمات مورد دسته‌بندی قرار می‌گیرند و آمار bigram و trigram بین دسته‌های کلمات استخراج می‌شود. ساده‌ترین روش برای دسته‌بندی کلمات، دسته‌بندی بر اساس پاره‌گفتار آن‌ها است. ازآنجاکه هر کلمه ممکن است بر اساس موقعیت و نقشی که در جمله دارد پاره‌گفتارهای مختلفی بگیرد، پس هر کلمه عضو چند طبقه خواهد بود. در پیکره متنی زبان فارسی برچسب کلمات، بیان‌گر پاره‌گفتار آن‌ها است. اگر T_i برچسب متناظر با کلمه

W_i باشد، مدل‌های bigram و trigram مبتنی بر پاره‌گفتار به‌ترتیب با روابط زیر به‌دست می‌آیند [7].

$$P(w_2|w_1) = P(T_2|T_1), P(w_2|T_2) = \frac{N(T_1T_2)}{N(T_1)}, \frac{N(w_2,T_2)}{N(T_2)} \quad (۱۲)$$

$$P(w_3|w_1w_2) = P(T_3|T_1T_2), P(w_3|T_3) = \frac{N(T_1T_2T_3)}{N(T_1T_2)}, \frac{N(w_3,T_3)}{N(T_3)} \quad (۱۳)$$

در این روابط $N(T_1)$، $N(T_1T_2)$ و $N(T_1T_2T_3)$ به‌ترتیب آمارهای monogram، bigram و trigram برچسب‌ها و $N(w_i.T_i)$ آمار lexical generation برای کلمه w_i و برچسب T_i (تعداد دفعاتی که کلمه w_i با برچسب T_i آمده است [7].

۵-۳- هموارسازی احتمالات bigram و trigram

وجود احتمال‌های صفر در مدل $n-gram$ محاسبات را درعمل با مشکل مواجه می‌کند، زیرا باعث می‌شود احتمال بسیاری از جمله‌های جدید برابر با صفر شود. از طرف دیگر این امر، در سامانهٔ بازشناسی نوشتار نیز باعث ایجاد مشکلاتی می‌شود؛ زیرا از این احتمالات، در هنگام بازشناسی لگاریتم گرفته می‌شود و لگاریتم عدد صفر مبهم است. برای حل این مشکل روش‌هایی موسوم به «هموارسازی» وجود دارند که سعی می‌کنند احتمال رخدادهای دیده‌نشده را به‌نحوی تخمین بزنند.

ساده‌ترین روش هموارسازی روش Add-One است. در این روش به همهٔ آمارهای bigram و trigram عدد ۱ اضافه می‌شود. به‌این‌ترتیب آمارهای صفر تبدیل به ۱ و درنتیجه به‌جای احتمالات صفر، احتمال کوچکی جایگزین می‌شود. احتمال‌های bigram و trigram پس از هموارسازی به روش زیر محاسبه می‌شوند:

$$P(w_2|w_1) = \frac{N(w_1w_2)+1}{N(w_1)+V} \quad (۱۴)$$

$$P(w_3|w_1w_2) = \frac{N(w_1w_2w_3)+1}{N(w_1w_2)+V} \quad (۱۵)$$

روش Add-One روش مؤثری نیست چون تغییر زیادی در احتمالات غیر صفر به وجود می‌آورد. بنابراین روش‌های هموارسازی دیگری مطرح‌شده‌اند، مانند روش هموارسازی خطی [24,25,33]. در بیشتر این روش‌ها برای تخمین احتمال‌های صفر، از احتمال‌های غیر صفر کمک گرفته می‌شود. بدین صورت که از احتمال‌های غیر صفر، توسط یکی از روش‌های «تخفیف»، کاسته شده و بر روی احتمال‌های صفر توزیع می‌شود [7].

در این پژوهش از روش هموارسازی «مطلق» استفاده شده است. روش‌های تخفیف به‌صورت کلی از رابطه زیر استفاده می‌کنند:

$$r^* = r, d_r \quad (۱۶)$$

که در آن r شمارش اولیه، r^* شمارش تخفیف داده‌شده و d_r ضریب تخفیف است.

روش «مطلق» مقدار d_r را با استفاده از رابطهٔ زیر تعیین می‌کند.

$$d_r = \frac{r-b}{r} \qquad b = \frac{n_1}{n_1+2n_2} \quad (۱۷)$$

که در آن n_1 تعداد شمارش‌های با مقدار ۱ و n_2 تعداد شمارش‌های با مقدار ۲ است.

با استفاده از رابطه‌های بالا از مقادیر غیر صفر کاسته می‌شود و باید این مقادیر بر روی شمارش‌های صفر به صورتی توزیع شود که مجموع احتمالات $n-gram$ هر سطر برابر با ۱ گردد.

بنابراین با فرض اینکه $P(w_i|w_{i-1}) = 0$ باشد، تخمین جدید $P(w_i|w_{i-1})$ به روش زیر به دست می‌آید:

$$\hat{P}(w_i|w_{i-1}) = \frac{\hat{\beta}(w_{i-1})}{\sum_{w_j:N(w_{i-1}w_j)=0} P(w_j)} P(w_i) \quad (۱۸)$$

$\hat{\beta}(w_{i-1})$ مجموع احتمالات کاسته‌شده در مرحلهٔ تخفیف است.

۶- به‌کارگیری مدل زبانی

در این بخش برای افزایش دقت بازشناسی، از مدل زبانی تولیدشده استفاده می‌گردد. معمولاً به دو صورت می‌توان از مدل زبانی استفاده کرد: «در حین جستجو» و «در انتهای جستجو» [27]. در نوشتار حاضر از روش در حین جستجو استفاده شده است.

به‌کارگیری مدل $n-gram$ در حین جستجو به این صورت است که هنگامی‌که الگوریتم جستجو فرضیه‌های مختلف را برای بازشناسی کلمات به جلو می‌برد، با شناسایی یک کلمه جدید، احتمال $n-gram$ آن را نیز همراه با امتیاز بازشناسی آن در امتیاز فرضیه ضرب می‌کند. بدین معنی که اگر امتیاز کنونی یک فرضیه پس از شناسایی کلمه w_n، S_n باشد و این فرضیه پس از بسط‌داده‌شدن، کلمه

w_{n+1} را به‌عنوان کلمه بعدی شناسایی کند، امتیاز جدید فرضیه برابر است با [7].

$$S_{n+1} = S_n . S_{HW}(w_{n+1}), S_{LM}(w_{n+1})^{LMW} \quad (۱۹)$$

که $S_{HW}(w_{n+1})$ امتیاز بازشناسی دست‌نوشته کلمه w_{n+1} و $S_{LM}(w_{n+1})$ امتیاز مدل زبانی آن است. به‌طورمعمول به‌دلیل تفاوت در مقیاس‌های $S_{HW}(w_{n+1})$ و $S_{LM}(w_{n+1})$، یک پارامتر وزن (LMW) برای امتیاز مدل زبانی در نظر گرفته می‌شود[16]. به‌طورمعمول برای پرهیز از به کار بردن اعداد خیلی کوچک به‌جای خود امتیازها از لگاریتم آن‌ها استفاده می‌شود[7].

$$\log S_{n+1} = \log S_{HW}(w_{n+1}) + LMW . \log S_{LM}(w_{n+1}) \quad (۲۰)$$

امتیاز نهایی تمامی فرضیه‌ها به همین روش با استفاده از ترکیب امتیازهای زبانی و بازشناسی نوشتار به‌دست می‌آید و درنهایت، فرضیه‌ای با بالاترین امتیاز، خروجی بخش بازشناسی سامانه خواهد بود. در این نوع از کاربرد مدل زبانی، رویه جستجو توسط مدل زبانی برای یافتن دنباله کلمات محتمل‌تر، هدایت می‌کند[7].

۷- هنجارسازی امتیاز بازشناسی

در این مقاله، برای محاسبه امتیاز بازشناسی، از روش نزدیک‌ترین همسایگی استفاده شده است [10]. در این روش ابتدا علائم کلمه ورودی بازشناسی شده و بر اساس آن‌ها فضای فرضیات محدود می‌شوند. در فضای فرضیات، همسایگی با استفاده از ویژگی زاویه و مختصات هنجارسازی‌شده مربوط به بدنهٔ زیرکلمات تعریف می‌شود و بر اساس این همسایگی به هر فرضیه یک امتیاز تعلق می‌گیرد؛ لذا در این روش فرضیه‌ای که کمترین فاصله را داشته باشد، بهترین فرضیه است، از این رو نمی‌توان به‌طورمستقیم از امتیاز بازشناسی در رابطه‌های (۱۹) و (۲۰) استفاده کرد؛ لذا در این پژوهش رابطه زیر برای هنجارسازی امتیازها طراحی شده است:

$$S_{i_{normal}} = \frac{-(S_i) + \max\limits_{1\le j\le H} S_j + 1}{\max\limits_{1\le j\le H} S_j - \min\limits_{1\le j\le H} S_j + 1} \quad (۲۱)$$

H تعداد فرضیات است. امتیاز حاصله عددی بزرگ‌تر از صفر و کوچک‌تر یا مساوی یک است؛ لذا هیچ امتیازی برابر صفر نیست و امتیاز یک متعلق به محتمل‌ترین فرضیه است.

به‌این‌ترتیب امتیاز بخش بازشناسی هنجارسازی‌شده و می‌توان از آن در هر دو رابطه مذکور استفاده نمود.

۸- تولید فرضیات جدید

در این مقاله مجموعه واژگان زبان فارسی به دو دسته تقسیم شدند. دستهٔ نخست واژگانی هستند که تمامی زیرکلمات آن‌ها توسط پایگاه داده زیرکلمات دست‌نویس [12] پشتیبانی می‌شوند؛ حال آنکه این دسته از واژگان ۶۸/۲٪ از کل واژگان را تشکیل می‌دهند و فرضیاتی که در مرحله بازشناسی امتیازدهی می‌شوند، اعضای این دسته از واژگان هستند. دستهٔ دوم واژگانی هستند که توسط پایگاه داده مذکور پشتیبانی نمی‌شوند. بدیهی است که اگر سامانه بازشناسی این دسته از واژگان را پشتیبانی نکند، قادر به بازشناسی بیش از سی درصد از کلمات زبان نخواهد بود؛ لذا در این بخش راه‌کاری برای پشتیبانی این دسته از واژگان ارائه می‌شود.

۹- انتخاب واژگان هم علامت با ورودی

در مرحله بازشناسی علائم موجود در ورودی تشخیص و یک برچسب علائم تولید می‌شود [10]. در این مرحله نیز با استفاده از همان برچسب، واژگان هم علامت با واژهٔ ورودی انتخاب می‌شوند. (این واژگان از مجموعه واژگانی انتخاب می‌شوند که در مرحله بازشناسی پشتیبانی نمی‌شدند).

۱۰- امتیازدهی به فرضیات جدید

در رابطه‌های (۱۹) و (۲۰) امتیازدهی به فرضیات، با تلفیق امتیازهای بازشناسی و مدل زبانی صورت می‌گرفت. آنچه مسلم است در این بخش به‌دلیل عدم وجود نمونه دست‌نویس زیرکلمه‌های تشکیل‌دهنده فرضیات، محاسبه امتیاز بازشناسی امکان‌پذیر نیست؛ لذا از امتیاز واژگان پشتیبانی‌شده که در مراحل قبل بازشناسی شد استفاده می‌شود.

طبق بررسی‌های انجام‌شده، این نتیجه حاصل شد که چنانچه واژهٔ معادل ورودی عضوی از مجموعه واژگان پشتیبانی شده باشد، حتی درصورتی‌که نتیجه بازشناسی اشتباه باشد، در اغلب موارد واژه‌ی صحیح در چهار فرضیه نخست قرار دارد. به‌طورمعمول امتیاز چندین فرضیه نخست نزدیک به هم بوده و سایر فرضیات، با فرضیهٔ صحیح فاصله

زیادی دارند. در (جدول (۹) فرضیات مختلف و امتیاز آن‌ها برای دو کلمه «تهران» و «ایران» آمده است.

(جدول-۹): فرضیات و امتیازها برای کلمات «تهران» و «ایران»
(Tabel-9): Assumptions and concessions for the "تهران" and "ایران"

فرضیات برای کلمه «تهران»			فرضیات برای کلمه «ایران»		
فرضیه	امتیاز	امتیاز نرمال	فرضیه	امتیاز	امتیاز نرمال
تهران	1699.4	1	ایران	1198.9	1
توان	2178.7	0.6472	ایوان	1454.7	0.8941
تهرانی	2317.4	0.5452	ایرانی	1825.3	0.7407
ستوان	2375.1	0.5027	ایوانی	2081.1	0.6347
توانی	2796.7	0.1924	امیران	2178.5	0.5944
متحدان	3057.1	0.0007	اسیران	3613.2	0.0004

همان‌طور که مشاهده می‌شود، امتیاز چند فرضیه نخست تا حدودی به امتیاز بهترین فرضیه نزدیک است؛ اما با دورشدن از فرضیه نخست، هم واژگان بسیار متفاوت‌تر از ورودی و هم امتیازاتشان است. ازآنجایی‌که سامانه به‌صورت برخط عمل می‌کند، باید از محاسبات غیرضروری اجتناب شود؛ لذا درصورتی‌که تعداد فرضیات در بخش بازشناسی بیش از چهار فرضیه باشد، فقط برای چهار فرضیه‌ی نخست امتیاز مدل زبانی محاسبه می‌شود. برای محاسبه امتیاز بازشناسی برای فرضیات جدید، درصورتی‌که در بخش بازشناسی کمتر از چهار فرضیه موجود باشد، کم‌ترین امتیاز فرضیه‌ها و در غیر این صورت امتیاز فرضیهٔ چهارم برای امتیاز بازشناسی فرضیات جدید در نظر گرفته می‌شود؛ سپس همانند فرضیات قبل برای فرضیات جدید نیز امتیاز زبانی را محاسبه و امتیاز نهایی برای هر فرضیه حاصل می‌شود.

۱۱- بازشناسی نهایی

درنهایت فرضیه‌ای که دارای بیش‌ترین امتیاز است، به‌عنوان خروجی سامانه در نظر گرفته و بقیه فرضیات نیز در خروجی به کاربر نمایش داده می‌شود. آزمایش‌ها نشان می‌دهد که حتی درصورت تشخیص اشتباه، واژهٔ صحیح در بیش‌تر مواقع به‌عنوان فرضیهٔ دوم و در برخی مواقع به‌عنوان فرضیهٔ سوم نمایش داده می‌شود.

۱۲- نتایج

ازآنجایی‌که دادهٔ استانداردی برای ارزیابی نتایج بازشناسی برخط فارسی در سطح کلمه یا جمله وجود ندارد، سه متن با موضوعات سیاسی، ورزشی و اقتصادی توسط یک نویسنده به‌صورت کلمه به کلمه نوشته‌شده است. نویسنده به قواعد سامانه آشنایی داشته و تمامی این قواعد را رعایت کرده است. نتایج بازشناسی بخشی از متن سیاسی در شکل‌های (۵و۶)، بخشی از متن ورزشی در شکل‌های (۷و۸) و بخشی از متن اقتصادی در شکل‌های (۹و۱۰) آمده است.

۱۲-۱- انواع خطاهای بازشناسی

در این مقاله خطاهای بازشناسی به سه دسته کلی تقسیم‌شده‌اند:

۱) **خطای جایگزینی:** این دسته خود به دو نوع خطا تقسیم می‌شوند:
نوع نخست: فرضیهٔ معادل ورودی در مجموعهٔ فرضیات وجود دارد؛ ولی فرضیهٔ دیگری به‌عنوان خروجی ارائه شده است.
نوع دوم: فرضیهٔ معادل ورودی در مجموعه فرضیات موجود نیست؛ لذا سامانه یک فرضیهٔ اشتباه ارائه کرده است.

۲) **خطای حذف:** این دسته از خطاها برای ورودی‌هایی رخ می‌دهد که هیچ فرضیه‌ای برای آن‌ها ارائه نشده است.

۳) **خطای درج:** این خطا فقط در هنگام بررسی نتایج در سطح حروف کاربرد دارد و به این معنی است که مثلاً کلمه ورودی دارای چهار حرف است؛ ولی کلمهٔ بازشناسی‌شده دارای شش حرف است.

در شکل (۵) خطاهای جایگزینی نوع اول با زیرخط، خطاهای جایگزینی نوع دوم با دو زیرخط و خطاهای حذف با علامت Ø نشان داده شده است.

(جدول-۱۰): نتایج بازشناسی متن سیاسی بدون مدل زبانی
(Tabel-10): Results of recognizing of a political text without using the language model

سطح بررسی	تعداد ورودی	خطای جایگزینی	خطای حذف	خطای درج	دقت
کلمات	220	16	7	-	89.6%
زیرکلمات	483	25	21	-	90.5%
حروف	1723	9	57	4	95.9%

نخستین گفت‌وگوی تلویزیونی رئیس دولت یازدهم بوای توضیح شرایط کشور و برنامه‌های دولت، این بار به میزبانی مجری جدید انجام شد. رئیس‌جمهور برنامه‌های کوتاه مدت دولت بوای مل مشکلات اقتصاد کشور و برخی مسائل سیاست خارجی را توضیح داد.

به گزارش تابناک، رئیس‌جمهور در آغاز این گفت‌وگو گفت: به همه مردم ایران، اقوام، روستانشین‌ها، شوی ها و همه آن‌هایی که الان در تلویزیون برنامه را Ø، سلام عرضی می‌کنم و درود می‌فرستم و هفته کراهتی، میلاد کریما اهل بیت Ø و میلاد امام شتم را که در پیش رو داریم، تبریک عرض می‌کنم و به همه خانواده‌های Ø در حادثه بسیار Ø دیشب تیت می‌گویم و بوای همه مجروحان این حادثه آرزوی سلامتی و بهبودی می‌کنم.

خدا را شاکریم در هفته‌های گذشته، تلاشی شده به وظایف دولت عمل شود؛ وزرا معرفی شوند. رأی اعتماد گرفتند و دولت تشکیل شو و به حمدالله سه هفت کاری از این دولت گذشته وØ خوشامد که به یکی از وعده‌هایی که به آن تأکید می‌کردم، عمل می‌کنم و آن گزارش دهی به مردم عزیز و بزرگوار کشورمان اتو. این دولت، دولت Ø خواهد بود و Ø بر مبنای تبلیغات نیست، بلکه برمبنای واقعیت‌ها و تشکیل واقعیت‌هاست.

(شکل-۵):بازشناسی بخشی‌از متن سیاسی بدون استفاده از مدل زبانی

(Figure-5): Recognizing a section of a political text without using the language model

نخستین گفت‌وگوی تلویزیونی رئیس دولت یازدهم برای توضیح شرایط کشور و برنامه‌های دولت، این بار به میزبانی مجری جدید انجام شد. رئیس‌جمهور برنامه‌های کوتاه مدت دولت برای حل مشکلات اقتصاد کشور و برخی مسائل سیاست خارجی را توضیح داد.

به گزارش تابناک، رئیس‌جمهور در آغاز این گفت‌وگو گفت: به همه مردم ایران، اقوام، روستانشین‌ها، شهری‌ها و همه آن‌هایی که الان در تلویزیون برنامه را می‌بینند، سلام عرض می‌کنم و درود می‌فرستم و هفته کرامت، میلاد کریمه اهل بیت علیه‌السلام و میلاد امام شتم را که در پیش رو داریم، تبریک عرض می‌کنم و به همه خانواده‌های جانباختگان در حادثه بسیار تأسف‌آور دیشب تیت می‌گویم و برای همه مجروحان این حادثه آرزوی سلامتی و بهبودی می‌کنم.

خدا را شاکریم در هفته‌های گذشته، تلاش شده به وظایف دولت عمل شود؛ وزرا معرفی شدند. رأی اعتماد گرفتند و دولت تشکیل شد و به حمدالله سه هفته کاری از این دولت گذشته و امشب خوشامد که به یکی از وعده‌هایی که به آن تأکید می‌کردم، عمل می‌کنم و آن گزارش دهی به مردم عزیز و بزرگوار کشورمان است. این دولت، دولت پاسخگو خواهد بود وØ بر مبنای تبلیغات نیست، بلکه برمبنای واقعیت‌ها و تشکیل واقعیت‌هاست.

(شکل-۶):بازشناسی بخشی از متن سیاسی با استفاده از مدل زبانی

(Figure-6): Recognizing a section of a political text using the language model

تساوی در پایان دربی ۷۷

استقلال و Ø در سومین دربی متوالی خود به تساوی بدون گل رضایت دادند.

به گزارش همشهری آنلاین، هفته شتم Ø برتر با بازی دو تیم بزرگ تهرانی به پایان رسید. در این بسازی که نقاد و Ø بازی دو تیم بود تماشاگران بار دیگری شاهد یک مسابقه بدون گل بودند.

سرخابی‌ها Ø دفاعی Ø بودند اما معلوم بود که اولویت اول آن‌ها گل نخوردن به هر Ø است. بازی با وضعیت تقریبا برابر در هر دو نیمه تداوم داشت و کمتر از Ø یک دست صحنه‌های Ø روی دروازه‌ها غلت شو. در این میان آنچه ثمود بود اصرار مدافعان دو تیم به دفع توپ‌ها بدون توجه به بازی‌سازی بود، مسئله‌ای که باعث شو به کرات Ø های دو تیم به عناصر بی‌عنایت این بازه تبدیل شوند. بسا این نتیجه استقلال ۱۵ امتیازی و Ø ۱۴ امتیازی همچنان در رده‌های چهارم و پنجم باقی ماندند.

(شکل-۷) :بازشناسی بخشی‌از متن ورزشی بدون استفاده از مدل زبانی

(Figure-7): Recognizing a section of a Sports text without using the language model

(جدول-۱۱): نتایج بازشناسی متن سیاسی با استفاده از مدل زبانی

(Tabel-11): Results of recognizing of a political text using the language model

سطح بررسی	تعداد ورودی	خطای جایگزینی	خطای حذف	خطای درج	دقت
کلمات	220	3	1	-	98.1%
زیرکلمات	483	4	4	-	98.3%
حروف	1723	2	10	-	99.3%

(جدول-۱۲) : نتایج بازشناسی متن ورزشی بدون مدل زبانی

(Tabel-12): Results of recognizing of a sports text without using the language model

سطح بررسی	تعداد ورودی	خطای جایگزینی	خطای حذف	خطای درج	دقت
کلمات	162	11	10	-	87%
زیرکلمات	344	18	15	-	90.4%
حروف	607	10	68	3	86.7%

(جدول-۱۳): نتایج بازشناسی متن ورزشی با استفاده از مدل زبانی

(Tabel-13): Results of recognizing of a sports text using the language model

سطح بررسی	تعداد ورودی	خطای جایگزینی	خطای حذف	خطای درج	دقت
کلمات	162	5	۰	-	96.9%
زیرکلمات	344	6	۰	-	98.2%
حروف	607	3	4	1	98.7%

(جدول-۱۴): نتایج بازشناسی متن اقتصادی بدون مدل زبانی
(Tabel-14): Results of recognizing of an economic text without using the language model

سطح بررسی	تعداد ورودی	خطای جایگزینی	خطای حذف	خطای درج	دقت
کلمات	117	13	2	-	87.2%
زیرکلمات	354	18	7	-	92.9%
حروف	475	14	18	5	92.2%

تساوی در پایان دربی ۷۷
استقلال و پرسپولیس در سومین دربی متوالی خود به تساوی بدون گل رضایت دادند.
به گزارش همشهری آنلاین، هفته شتم لیگ برتر با بازی دو تیم بزرگ تهرانی به پایان رسید. در این بازی که هفتاد و هفتمین بازی دو تیم بود تماشاگران بار دیگری شاهد یک مسابقه بدون گل بودند.
سرخابی‌ها ترکیب دفاعی نچیده بودند اما معلوم بود که اولویت اول آن‌ها گل نخوردن به هر قیمت است. بازی با وضعیت تقریبا برابر در هر دو نیمه تداوم داشت و کمتر از انگشتان یک دست صحنه‌های هیجان‌انگیز روی دروازه‌ها غلت شد. در این میان آنچه مشهود بود اصرار مدافعان دو تیم به دفع توپ‌ها بدون توجه به بازی‌سازی بود، مسئله‌ای که باعث شد به کرات هافبک های دو تیم به عناصر بی‌نهایت این بازی تبدیل شوند.
با این نتیجه استقلال ۱۵ امتیازی و پرسپولیس ۱۴ امتیازی همچنان در رده‌های چهارم و پنجم باقی ماندند.

(شکل-۸): بازشناسی بخشی از متن ورزشی با استفاده از مدل زبانی
(Figure-8): Recognizing a section of a Sports text using the language model

(جدول-۱۵): نتایج بازشناسی متن اقتصادی با استفاده از مدل زبانی
(Tabel-15): Results of recognizing of an economic text using the language model

سطح بررسی	تعداد ورودی	خطای جایگزینی	خطای حذف	خطای درج	دقت
کلمات	117	5	۰	-	95.7%
زیرکلمات	354	8	۰	-	97.7%
حروف	475	5	2	2	98.1%

معاملات ۲ نوع اوراق مشارکت شرکتی ملی نفت در بودی تهران آغاز شد طبق اعلام بودی اوراق بهادار ۵۰ میلیون برگه از اوراق مشارکتی ملی نفت به ارزشی ۵۰ هزار میلیارد ریال در بودی گشوده شد. این اوراق با سر رسید ۳ساله و نرخ سود Ø ۲۱درصد روزشمار و معاف از مالیات است که سود آن هر سه‌ماه پرداخت می‌شود. دانه نوسان روزانه اوراق ۵ درصد است ضامن اوراق مزبور صندوقی Ø پس‌انداز و رفاه کارکنان صنعتی نفت و عامل پرداختی سود سپرده‌گذاری مرکزی اوراق بهادار معرفی شده است. ۲۵ افضل پارسال، با عرضه و فروشی ۴۸هزار میلیارد ریال اوراق مشارکت ملی نفت با بزرگ‌ترین تمایز مالی دولت در تاریخی بودی رقم خورد.

(شکل-۹) : بازشناسی بخشی‌از متن اقتصادی بدون استفاده از مدل زبانی
(Figure-9): Recognizing a section of an economic text without using the language model

معاملات ۲ نوع اوراق مشارکت شرکتی ملی نفت در بودی تهران آغاز شد طبق اعلام بودی اوراق بهادار، ۵۰ میلیون برگه از اوراق مشارکت ملی نفت به ارزش ۵۰هزار میلیارد ریال در بورس گشوده شد. این اوراق با سر رسید ۳ساله و نرخ سود علی الحساب ۲۱درصد روزشمار و معاف از مالیات است که سود آن هر سه‌ماه پرداخت می‌شود. دانه نوسان روزانه اوراق ۵ درصد است ضامن اوراق مزبور صندوق بازنشستگی پس‌انداز و رفاه کارکنان صنعتی نفت و عامل پرداخت سود سپرده‌گذاری مرکزی اوراق بهادار معرفی شده است. ۲۵اسفندپارسال، با عرضه و فروش ۴۸هزار میلیارد ریال اوراق مشارکت ملی نفت با بزرگ‌ترین تمایز مالی دولت در تاریخ بورس رقم خورد.

(شکل-۱۰): بازشناسی بخشی از متن اقتصادی با استفاده از مدل زبانی
(Figure-10): Recognizing a section of an economic text using the language model

۱۲-۲- تحلیل نتایج

همان‌طور که مشاهده می‌شود، در شکل (۵) بخشی از متن سیاسی حاصل از بازشناسی بدون استفاده از مدل زبانی خطاهای زیادی رخ داده است، که بخش قابل توجهی از این خطاها، خطاهای حذف (در سطح کلمه) است. در روش پیشنهادی تدابیری برای افزودن فرضیات جدید ارائه شد که این نوع خطاها را تا حد بالایی برطرف می‌کند.

در جدول (۱۱) نتایج بازشناسی بر اساس مدل Bigram مبتنی بر کلمه آورده شده است. همان‌طور که مشاهده می‌شود تعداد خطاهای (در سطح کلمه) جایگزینی به سه خطا و تعداد خطاهای حذف به یک خطا کاهش‌یافته است. خطاهای جایگزینی در بازشناسی کلمات «هشتم»، «تسلیت» و «خوشحالم» روی داده است. دلیل بروز این خطاها این است که فرضیات «شتم»، «تیت» و «خوشامد» توسط پایگاه داده دست‌نویس پشتیبانی می‌شوند و لذا امتیاز بالایی در بخش نزدیک‌ترین همسایگی به آن‌ها تعلق می‌گیرد، از طرف دیگر کلمات «هشتم»، «تسلیت» و «خوشحالم» توسط پایگاه داده پشتیبانی نمی‌شود؛ لذا این فرضیات امتیاز بخش بازشناسی خود را از چهارمین فرضیه کسب می‌کنند، در بخش مدل زبانی نیز فرضیات امتیازی را کسب می‌کنند، ولی این امتیازها نمی‌تواند فرضیات درست را به محتمل‌ترین فرضیات تبدیل کند؛ چراکه امتیاز چهارمین فرضیه اختلاف بسیار زیادی با امتیاز فرضیات نخست دارد. خطای حذف نیز مربوط به کلمه «گزارشش» است. با تدابیری که برای افزودن فرضیات

جدید در الگوریتم پیشنهادی اعمال شد بسیاری از خطاهای حذف برطرف می‌شود؛ ولی کلمهٔ مذکور چه در واژگان قابل پشتیبانی توسط پایگاه‌داده دست‌نویس و چه در واژگان پشتیبانی‌نشده موجود نیست؛ لذا این خطا در تمامی بازشناسی‌های مختلف این متن رخ داده است.

بیش‌تر خطاهایی که در شکل (۷) اتفاق افتاده؛ از نوع خطای حذف یا خطای جایگزینی نوع دوم است. این نمونه خطاها نشان می‌دهد که پایگاه داده زیرکلمات دست‌نویس نمی‌تواند به‌خوبی کلمات و اصطلاحات ورزشی را پوشش دهد. نکتهٔ دیگری که در جدول مذکور قابل‌توجه است، بازشناسی کلمهٔ «شد» است. این کلمه چند بار به‌اشتباه «شو» بازشناسی شده است که علت این امر به دستخط نویسنده در نوشتن شکل خاصی از کلمهٔ «شد» برمی‌گردد. در شکل (۸) اعمال روش پیشنهادی و استفاده از مدل زبانی تعداد قابل توجهی از خطاها برطرف شده است و محدود خطاهایی که هنوز وجود دارند، خطای جایگزینی نوع نخست است. این نتایج نشان می‌دهد که با استفاده از مدل زبانی، در خروجی این متن هیچ خطای حذف یا خطای جایگزینی نوع دومی وجود ندارد؛ لذا تمامی خطاها، خطاهای جایگزینی نوع نخست بوده که به‌دلیل عدم پوشش بخشی از زیرکلمات ورودی توسط پایگاه داده دست‌نویس و اختلاف امتیاز بازشناسی چهار فرضیهٔ نخست، فرضیهٔ صحیح توان رقابت با فرضیه‌ای که موفق به کسب رتبهٔ نخست شده را ندارد. همان‌گونه که درقبل اشاره شد، بخشی از ضعف‌های تشخیص، مربوط به عدم پشتیبانی پایگاه داده زیرکلمات دست‌نویس از برخی کلمات با موضوع ورزشی می‌شود. برای بررسی بیشتر عملکرد الگوریتم در موضوعات مختلف، کارآیی آن بر روی یک متن اقتصادی نیز مورد بررسی قرار گرفت. نتایج در جدول‌های (۱۴ و ۱۵) نشان داده شده است. مشاهده می‌شود که نتایج با استفاده از مدل زبانی بهبود چشم‌گیری داشته است. البته برخی از کلمات مانند «بورس»، به‌دلیل اینکه توسط پایگاه داده پشتیبانی نمی‌شود، به‌درستی تشخیص داده نشده است. در این متن نیز مشاهده می‌شود که با استفاده از مدل زبانی، خطاهای حذف از بین رفته است. با بررسی نتایج سه متن سیاسی، ورزشی و اقتصادی مشاهده می شود که بخش تشخیص توسط مدل زبانی با توجه به اینکه پیکره متنی برخی موضوعات را بیشتر پوشش داده است، دقت بازشناسی در موضعات مختلف متفاوت است. به‌عنوان‌مثال در موضوعات سیاسی نسبت به موضوعات ورزشی مدل زبانی بهتری را می‌توان استخراج کرد. البته نتایج به‌دست‌آمده نشان‌دهندهٔ بهبود عملکرد الگوریتم پیشنهادی در هر سه موضوع است و تفاوت چندان زیادی در عملکرد آن در موضوعات مختلف مشاهده نمی‌شود.

۱۲-۳- مقایسه با سایر روش‌ها

برای مقایسه روش پیشنهادی با سایر روش‌ها، تاکنون کار مشابهی که در آن از مدل زبانی در تشخیص دست‌نوشته برخط استفاده شده باشد، ارائه نشده است که بتوان مقایسه انجام داد؛ ولی نزدیک‌ترین مقاله به نوشتار پیش رو، مقاله [10] است که به‌کمک یک واژگان گسترده از کلمات فارسی فرایند بازشناسی را که تا قبل از آن در سطح زیر کلمه انجام می‌شد، در سطح کلمه موردبررسی قرار داده، که نتایج به‌دست‌آمده در سطح کلمه، زیرکلمه و حروف به‌ترتیب۸۷/۳ ۹۳/۲ و ۹۴/۹ درصد گزارش شده است. شکل (۷) نتیجه پیاده‌سازی مقاله مذکور است (البته با مجموعه اصلاحاتی که در بخش ۲ نوشتار حاضر ارائه شده است)، همان‌طور که مشاهده می‌شود، در این آزمایش کلماتی مانند: «پرسپولیس»، «لیگ»، «نچیده»، «هفتمین»، «انگشتان» و... شناسایی نشده‌اند و به‌اصطلاح خطای حذف رخ داده است، عامل اصلی رخ دادن این خطا، عدم پشتیبانی پایگاه داده زیرکلمات دست‌نویس از زیرکلمات کلمات مذکور است، لذا در این روش به‌دلیل اینکه کلمه موردنظر در فرضیات موجود نیست، امکان بازشناسی این کلماتی غیرممکن است. در شکل (۸) مشاهده می‌شود که با اعمال مدل زبانی خطاهای از نوع حذف اتفاق نیفتاده است و با اینکه پایگاه داده دست‌نویس از زیرکلمات مذکور پشتیبانی نمی‌کند، این کلمات با موفقیت شناسایی شده‌اند.

در ادامه روش‌های ارائه‌شده در [5]،[4] به‌وسیله برنامه Matlab R2012a شبیه‌سازی شد، که نتایج مشاهده‌شده به‌شرح زیر است (هر دو روش برای بازشناسی زیرکلمات ارائه‌شده‌اند): در [4] بازشناسی با استفاده از توابع عضویت فازی و مدل تطبیق رشته انجام‌شده است، و [5] یک سامانه عصبی فازی ارائه شده و نتایج به‌دست‌آمده در جدول (۱۷) قابل‌مشاهده است:

(جدول-۱۶): بازشناسی متن ورزشی [4]
(Tabel-16): Recognizing of a sports text [4]

سطح بررسی	تعداد ورودی	خطای جایگزینی	خطای حذف	خطای درج	دقت
زیرکلمات	344	35	۰	-	87.5%
حروف	607	18	104	4	79.2%

(جدول-۱۷): بازشناسی متن ورزشی [5]
(Tabel-17): Recognizing of a sports text [5]

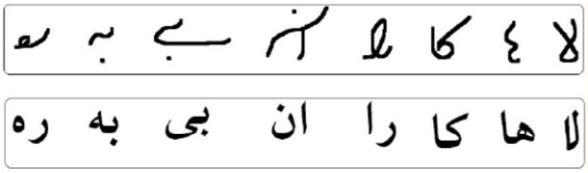
سطح بررسی	تعداد ورودی	خطای جایگزینی	خطای حذف	خطای درج	دقت
زیرکلمات	344	35	۰	-	90%
حروف	607	20	99	3	79.9%

با آزمایش ورودی‌های مختلف در هر دو روش به این نتیجه رسیدیم که اگر زیرکلمات تشکیل‌دهنده کلمه واردشده توسط پایگاه داده نمونه‌های دست‌نویس پشتیبانی شود نتیجه بازشناسی زیرکلمه مطلوب است؛ ولی درصورت عدم پشتیبانی زیرکلمه به‌دلیل اینکه زیرکلمه مورد نظر در فرضیات موجود نیست، نتیجه به‌یقین با خطای جایگزینی مواجهه می‌شود. درصورتی‌که در روش پیشنهادی نوشتار پیش رو، مجموعه واژگان زبان فارسی به دو دسته تقسیم شدند. دستهٔ نخست واژگانی هستند که تمامی زیرکلمات آن‌ها توسط پایگاه داده زیرکلمات دست‌نویس پشتیبانی می‌شوند، این دسته از واژگان ۶۸/۲٪ از کل واژگان را تشکیل می‌دهند (که البته اغلب واژگان پرکاربرد را در این دسته قرار دارند) و دستهٔ دوم واژگانی هستند. که توسط پایگاه داده مذکور پشتیبانی نمی‌شوند. روش‌هایی که تاکنون ارائه شده‌اند (از جمله هر دو روشی که برای این آزمایش انتخاب شده‌اند) قادر به پشتیبانی واژگان دسته دوم نیستند؛ بااین‌حال در روش پیشنهادی به کمک مدل زبانی این کلمات نیز قابل‌بازشناسی است. جدول (۱۳) گویای این است که استفاده از مدل زبانی در فرآیند بازشناسی دست‌نوشته‌های برخط می‌تواند به‌طور چشمگیری مؤثر واقع شود.

۱۲-۴- نتیجه‌گیری

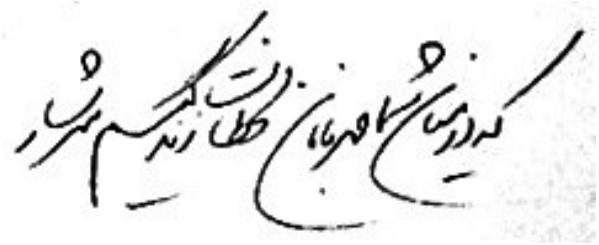
با بررسی نتایج ارائه‌شده در جدول (۱۴) مشاهده می‌شود که با استفاده از این روش برای بازشناسی یک متن سیاسی، دقت بازشناسی در سطح حروف از ۹۵/۹٪ به ۹۹/۳٪ ارتقا یافته است؛ لذا نتایج حاصل از ارزیابی روش پیشنهادی نشان می‌دهد که استفاده از مدل زبانی در سامانهٔ بازشناسی برخط نوشتار می‌تواند باعث بهبود فرآیند بازشناسی شوند.

استفاده از روش پیشنهادی، نقاط ضعفی نیز دارد که امید است در آینده و ادامه پژوهش‌ها در این حوزه مرتفع شود، به‌عنوان‌مثال قواعدی که کاربر در ورود اطلاعات ناچار به رعایت آن‌هاست. هرچند در این مقاله سعی شده است نسبت به روش‌های پیشین قواعد دست و پاگیر تا حد امکان کمتر شود، ولی این موضوع همچنان می‌تواند به‌عنوان یک نقطه‌ضعف محسوب شود. شاید بتوان عدم توانایی تشخیص حروف ادغام‌شده را به‌عنوان بزرگ‌ترین ضعف الگوریتم پیشنهادی مطرح کرد. در نوشته‌های دست‌نویس فارسی به‌خاطر زیبایی بصری نوشتار و همچنین سلیقه نویسنده، شکل بعضی از حروف کنار هم، به‌کلی تغییر می‌کند که به آن‌ها حروف ادغام‌شده می‌گویند. ازآنجایی‌که تولید فرضیات الگوریتم پیشنهادی به‌طورکامل وابسته به تعداد حرکات قلم است، در هنگام واردکردن حروف ادغام‌شده، در بیش‌تر مواقع به‌دلیل عدم مطابقت تعداد حرکت قلم با الگوی اصلی، واژه صحیح در میان فرضیات قرار ندارد. نمونه‌ای از این واژه‌ها در شکل (۱۱) قابل‌مشاهده است.



(شکل-۱۱) : نمونه‌ای از حروف ادغام‌شده در فارسی
(Figure-11): A sample of the Persian merged letters

همچنین در مواردی که زیرکلمات وارد شده هم‌پوشانی زیادی با یکدیگر داشته باشند، خطای برنامه افزایش می‌یابد. نمونه‌هایی از آن در شکل (۱۲) آمده است.



(شکل-۱۲) : نمونه‌ای از کلماتی با هم‌پوشانی فراوان
(Figure-12): A sample of very overlapping words

علاوه بر این، ازآنجایی‌که مدل‌های زبانی به‌شدت تحت تأثیر ویژگی‌های لغوی، نحوی و معنای آموزش دیده است، نبود دادگان مناسب در این زمینه نیز از موانع پژوهش در این حوزه محسوب می‌شود. یکی دیگر از مشکلاتی که در این حوزه می‌توان به آن اشاره کرد، عدم وجود و پایگاه داده دست‌نویس مناسب است.

به‌عنوان کارهای پژوهشی آینده، علاوه‌بر رفع موانع

مطرح‌شده، پیشنهاد می‌شود، مدل‌هایی برای بازشناسی تولید شوند که توانایی تطبیق‌پذیری با نویسنده را داشته باشند. برای این کار می‌توان از الگوریتم‌های «یادگیری تقویتی» استفاده کرد.

۱۳- مراجع 13-References

[1] T. Ebrahimi and B. Z. Dehkordi," Using of factor oriented language model for increase of speech recognition rate," 1st conference on new ideas in computer engineering, Sharekord,2015.

[2] نوش‌آبادی فخری، احمدی فرد علیرضا، خسروی حسین، "تشخیص دست‌نوشته برخط فارسی با رویکرد تجزیه‌ای"، پایان‌نامه کارشناسی ارشد؛ دانشگاه صنعتی شاهرود؛ ۱۳۹۳.

[2] F. Nooshabadi, A. Ahmadifard, H. Khosravi," Online persian hand writing recognition using Analytical approach", M.S. Thesis, Shahrood university of technology, 2014.

[3] مهرعلیان محمدامین و فولادی کاظم، "بازشناسی برخط حروف مجزای دست‌نویس فارسی بر اساس تشخیص گروه و بدنه اصلی با استفاده از ماشین بردار پشتیبان"، هفتمین کنفرانس ماشین بینایی و پردازش تصویر، تهران، دانشگاه علم و صنعت، ۱۳۹۰.

[3] M. Mehralian and K. Fooladi," Online persian hand writing discrete letter recognition based on group and main body detection using SVM", 7th Iranian conference on machine vision and image processing, Tehran, 2011.

[4] اسمعیل‌پور ندا، برومندنیا ندا، "بازشناسی زیر-کلمات برخط فارسی بر اساس رویکرد فازی و ساختاری با استفاده از ساختار لیست‌های پیوندی"، یازدهمین کنفرانس سراسری سیستم‌های هوشمند. انجمن سیستم‌های هوشمند ایران، ۱۳۹۱.

[4] N. Esmailpour and N. Broomandnia, "Recognition of Persian online sub-words based on fuzzy and structural approach using link list structure", 11th Iranian conference on intelligent systems, 2012.

[5] خوش کلام محصصی زهرا، رضوی ابراهیمی سید علی و فرهودی‌نژاد اکبر، «طراحی یک سیستم عصبی- فازی با قابلیت آموزش هم‌زمان برای بازشناسی بر خط زیر - کلمات فارسی»، همایش ملی مهندسی کامپیوتر و توسعه پایدار با محوریت شبکه‌های کامپیوتری، مدل‌سازی و امنیت سیستم‌ها، مشهد، موسسه آموزش عالی خاوران، ۱۳۹۲.

[5] Z. K. Mohassesi, S.A. Ebrahimi and A. Farhoodinezhad," The design of a neuro-fuzzy system with simultaneous training for on line recognizing the Persian sub-words", 8th Symposium on advances in science and technology (computer networks, modelling and system security), Mashahd, 2013.

[6] امینیان، مریم، "خوشه‌بندی معنایی افعال زبان فارسی"، پایان‌نامه کارشناسی ارشد؛ دانشگاه صنعتی شریف؛ ۱۳۹۱.

[6] M. Imanian, "Semantic clustering of verbs in Persian language", M.S. thesis, Sharif university of technology, 2012.

[7] بحرانی محمد، ثامتی حسین، حافظی نازیلا، ممتازی سعیده، موثق حامد، "به‌کارگیری پیکره متنی زبان فارسی در ساخت مدل‌های زبانی آماری برای سیستم‌های بازشناسی گفتار پیوسته فارسی"، دومین کارگاه پژوهشی زبان فارسی و رایانه، صص ۹۲-۱۰۹، ۱۳۸۵.

[7] M. Bahrani, H. Sameti, N. Hafezi, S. Momtazi and H. Movasegh, "The use of the Persian text framework in the production of statistical language models for Persian continuous speech recognition systems", 2nd workshop of Persian language and computer, PP. 92-109, 2006.

[8] پیرنیا نائینی، شهریار و خادمی، مریم، "قطعه‌بندی برخط دست‌نویس فارسی با استفاده از استخراج ویژگی‌ها" ، سومین همایش ملی کامپیوتر و فناوری اطلاعات مهندسی سما، صص ۲۶۶-۲۷۱، همدان، ایران، ۱۳۸۹.

[8] Sh. P. Naeini and M. Khademi, "Persian online handwriting fragmentation using feature extraction", 3rd national conference on computer engineering and information technology, P.P. 266-271, Hamedan, Iran, 2010.

[9] رضوی، سید محمد و کبیر، احسان‌الله، "بازشناسی برخط حروف مجزای فارسی با شبکه عصبی"، سومین کنفرانس ماشین بینایی و پردازش تصویر، جلد ۴۱ شماره ۱، صص ۸۳-۸۹، دانشگاه تهران،۱۳۸۳.

[9] S. M. Razavi and E. Kabir, “Online Persian hand writing discrete letter recognition using neural network”, 3rd Iranian conference on machine vision and image processing, P.P. 83-89, Tehran, Iran, 2004.

[10] رضوی، سید محمد و کبیر، احسان‌الله، "بازشناسی برخط کلمات دست‌نویس فارسی با واژگانی گسترده"، ۱۳۸۷، پنجمین کنفرانس ماشین بینایی و پردازش تصویر.

[10] S. M. Razavi and E. Kabir, “Online Persian hand writing words recognition by Extensive vocabulary”, 5th Iranian conference on machine vision and image processing, Iran, 2008.

[11] رضوی، سید محمد و کبیر، احسان‌الله، "روشی ساده برای بازشناسی برخط زیر-کلمات فارسی"، نشریه مهندسی برق و مهندسی کامپیوتر ایران، شماره ۲، صص ۶۳- ۷۲، ۱۳۸۴.

[11] S. M. Razavi and E. Kabir, “A simple way to recognize online Persian sub-words”, Iranian journal of electrical and computer engineering, vol. 2, P.P. 63-72, 2005.

[12] رضوی، سید محمد و کبیر، احسان‌الله، "یک پایگاه داده برای بازشناسی دست‌نوشته‌های برخط فارسی"، ششمین کنفرانس سیستم‌های هوشمند، کرمان، ۱۳۸۳.

[12] S. M. Razavi and E. Kabir, “A database for recognizing Persian online hand writing”, Iranian journal of electrical and computer engineering, 6th Iranian conference on intelligent systems, Kerman, Iran, 2004.

[13] ساجدی، هدیه و جمزاده، منصور و ثامتی، حسین و باباعلی، باقر، "ارائه‌ی یک روش مبتنی بر گروه‌بندی برای بازشناسی حروف مجزای برخط فارسی به کمک مدل مخفی مارکوف "، دوازدهمین کنفرانس بین‌المللی انجمن کامپیوتر ایران، صص ۴۱۹-۴۲۵، دانشگاه تهران، ۱۳۸۵.

[13] H. Sajedi, M. Jamzadeh, H. Sameti, B. Babaali “Presentation of a Grouping-Based Approach to Recognition of Persian Separated Letters Using the Hidden Markov Model”, 12th International conference of Iranian computer society, Tehran, Iran, 2006.

[14] قدس، وحید و کبیر، احسان‌الله، "بررسی شیوه‌های متداول نگارش دست‌نوشته‌های برخط فارسی به‌منظور استفاده در بازشناسی آن‌ها"، مجله مهندسی برق دانشگاه تبریز، جلد ۴۱ شماره ۱، صص ۲۲-۳۲، ۱۳۹۱.

[14] V. Ghods and E. Kabir, “The study of common ways of Persian online hand writing for use in their recognition”, Tabriz journal of electrical engineering, Vol. 41, No. 1, P.P. 22-32, 2012.

[15] فرهنگستان زبان و ادب فارسی (نشر آثار) ، دستور خط فارسی، چاپ نهم، ۱۳۸۹.

[15] “Persian Language and Literature Academy (Publishing Works)”, Persian writing order, 9th edition, 2010.

[16] کبودیان جهانشاه، شجاع مودب حمیدرضا، شیخ زادگان جواد، "یک سیستم جستجوگر کلمات مبتنی بر مدل پنهان مارکوف با دایره لغات نامحدود برای جستجوی مستندات گفتاری در محیط‌های واقعی و عملیاتی"، دهمین کنفرانس سالانه انجمن کامپیوتر ایران، ۱۳۸۳.

[16] J. Kaboodian, H. S. Moadab and J. Shaikhzadegan, “A word search engine based on the hidden Markov model with unlimited vocabulary to search for spoken documentation in real-world environments”, 10th national conference of Iranian computer society, 2004.

[17] میرزازاده، فرزانه، "بازشناسی کلمات در دست‌نوشته بر خط فارسی به روش فازی"، پایان‌نامه کارشناسی ارشد، دانشگاه صنعتی شریف، تهران، ۱۳۸۶.

[17] F. Mirzadeh, “Fuzzy based recognition of words in the Persian hand writing”, M.S. Thesis, Sharif university of technology, 2007.

[18] پژوهشنامه نویسه‌خوان نوری OCR فارسی، شورای پژوهشی OCR کارگروه خط و زبان فارسی شورای عالی اطلاع‌رسانی, پائیز ۱۳۸۶.

[18] “Research papers of Persian OCR”, The OCR Research Council of the Persian writing and Language Teams, 2007.

[19] همایون پور محمد مهدی، سلیمی بدر آرمین، "تعیین مرز و نوع عبارات نحوی در متون فارسی"، فصلنامه علمی-پژوهشی پردازش علائم و داده ها، جلد ۱۰، شماره ۲، صفحه ۶۹-۸۶، ۱۳۹۲.

[19] M. M. Homayoonpor and A. Salimi Badr, "Determining the boundary and type of syntactic expressions in Persian texts", Signal and data processing, Vol. 10, No. 2, P.P. 69-86, 2013.

[20] بایسته تاشک الهام، احمدی فرد علیرضا، خسروی حسین، " روشی دو مرحله ای برای بازشناسی کلمات دست نوشته فارسی به کمک بلوک بندی تطبیقی گرادیان تصویر"، فصلنامه علمی-پژوهشی پردازش علائم و داده ها، جلد ۱۲، شماره ۳، صفحه ۱۵-۲۹، ۱۳۹۴.

[20] E. B. Tashk, A. Ahmadifard and H. Khosravi, "A two-step method for recognizing Persian handwritten words using the adaptive blocking of image gradients", Signal and data processing, Vol. 12, No. 3, P.P. 15-29, 2015.

[21] دیانت روح الله، علی احمدی مرتضی، اخلاقی محمد یحیی، باباعلی باقر، " ارایه یک روش جدید بازیابی اطلاعات مناسب برای متون حاصل از بازشناسی گفتار"، فصلنامه علمی-پژوهشی پردازش علائم و داده ها، جلد ۱۳، شماره ۴، صفحه ۹۳-۱۰۸، ۱۳۹۵.

[21] R. Deinat, M. Aliahmadi, M. Y. Akhlaghipour and B. Babaali, "Introducing a new information retrieval method applicable for speech recognized texts", Signal and data processing, Vol. 13, No. 4, P.P. 93-108, 2016.

[22] C. L. Liu, S. Jaeger, and M. Nakagawa, "Online recognition of Chinese characters: the state-of-the-art", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 2, pp. 198–213, 2004.

[23] D. Jurafsky and J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson Prentice Hall, 2009.

[24] H. N. Eseen and R. Kneser, "On Structuring Probabilistic Dependencies in Stochastic Language Modeling", Computer, Speech, and Language, vol. 8, pp. 1–38, 1994.

[25] H. Witten and T. C. Bell, "The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression", IEEE Transactions on Information Theory, vol. 37, no. 4, pp. 1085–1094, 1991.

[26] M. Bijankhan, "The role of the corpus in writing a grammar: An introduction to a software", Iranian Journal of Linguistics, vol. 19, no. 2, 2004.

[27] M. P. Harper, L. H. Jamieson, C. D. Mitchell, G. Ying, S. Potisuk, P. N. Srinivasan, R. Chen, C. B. Zoltowski, L. L. McPheters, and B. Pellom, "Integrating language models with speech recognition". AAAI-94 Workshop on the Integration of Natural Language and Speech Processing, Seattle, Washington, pp. 139-146, 1994.

[28] R. Plamondon and S. Srihari, "Online and off-line handwriting recognition: a comprehensive survey", Pattern Analysis and Machine, vol. 22, no. 1, pp. 63–84, 2000.

[29] S. Al-Emami and M. Usher, "On-line recognition of handwritten Arabic characters", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 7, pp. 704–710, 1990.

[30] S. Atkins, J. Clear, and N. Ostler, "Corpus design criteria", Literary and linguistic computing, vol. 7, no. 1, pp. 1–16, 1992.

[31] S. Connell and A. Jain, "Online handwriting recognition using multiple pattern class models", Michigan State University, 2000.

[32] S. Jaeger, C. L. Liu, and M. Nakagawa, "The state of the art in Japanese online handwriting recognition compared to techniques in western handwriting recognition", International Journal on Document Analysis and Recognition, vol. 6, no. 2, pp. 75–88, 2003.

[33] S. Katz, "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, no. 3, pp. 400–401, 1987.

سلمان مسکنتی مدارک کاردانی و کارشناسی خود را در رشته کامپیوتر گرایش نرم‌افزار به‌ترتیب از دانشگاه شهید چمران اهواز و دانشگاه آزاد واحد بهبهان در سال‌های ۱۳۸۷ و ۱۳۸۹دریافت کرد. ایشان مدرک کارشناسی ارشد خود را در رشته هوش‌مصنوعی از دانشگاه علوم و تحقیقات بوشهر در سال ۱۳۹۲ اخذ کرد. ایشان از سال ۱۳۹۴ به‌عنوان پژوهش‌گر با آزمایشگاه ماشین بینایی و هوش مصنوعی دانشگاه خلیج فارس همکاری دارد. زمینه‌های پژوهشی مورد علاقه ایشان توسعه سامانه‌های Enterprise ، پژوهش در زمینه امنیت نرم‌افزار، پژوهش بر روی فناوری‌های نوین در حوزه برنامه‌نویسی و هوش مصنوعی می باشد.

نشانی رایانامه ایشان عبارت است از:

Maskanati@gmail.com

احـــمد کشاورز مدارک کارشناسی و کارشناسی ارشد خـــود را به‌ترتیب در سال‌های ۱۳۸۰ و ۱۳۸۳ از دانشـــگاه شیراز و دانشگاه تربیت مدرس در رشته مهندسی برق و مـــخابرات- سیســتم دریافت کرد. ایشان درجه دکترای خود را در ســـال ۱۳۸۷ از دانشگاه تربیت مدرس در رشته مخابرات سیستم دریافت کرد. وی هم اکنون استادیار گروه مهندسی بـــرق دانشگاه خلیج فارس است. زمینه‌های پژوهشی مورد علاقه ایشان عبارت است از: سنجش از دور، پردازش تصاویر پزشکی، ماشین بینایی و هوش مصنوعی.

نشانی رایانامه ایشان عبارت است از:

A.keshavarz@pgu.ac.ir