طراحی و پیاده‌سازی سامانهٔ شناسایی و تصحیح خطای املایی متون فارسی مبتنی بر معنای واژگان

محمدباقر دستغیب[1*]، سارا کلینی[2] و سید مصطفی فخر احمد[3]

[1] گروه پژوهشی طراحی و عملیات سیستم‌ها، مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری

[2] هوش ماشین و رباتیک، مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری

[3] بخش علوم و مهندسی کامپیوتر، دانشکده برق و مهندسی کامپیوتر، دانشگاه شیراز، شیراز، ایران

چکیده

طراحی و پیاده‌سازی ابزارهای پردازش زبان طبیعی فارسی، بر اساس ویژگی‌های خاص این زبان، همواره با چالش‌هایی مواجه است. با توجه به این‌که سامانه‌های تصحیح املای خودکار در حوزه‌های مختلفی از قبیل تصحیح پرس‌وجوها، بررسی املای واژگان در اینترنت و برنامه‌های ویراستاری متنی کاربرد دارد، لازم است تا برای زبان فارسی نیز نرم‌افزارهای مناسب ایجاد شود. در این مقاله ابتدا مقدمه‌ای درخصوص انواع خطاهای املایی، راه‌کارهای شناسایی و تصحیح خطاها شرح داده شده و سپس به معرفی سامانه پارسی‌اسپل که بر اساس معنای واژگان فارسی، خطاها را شناسایی و تصحیح می‌کند، می‌پردازیم. با توجه به نتایج حاصله از ارزیابی سامانه پارسی‌اسپل با سایر نرم‌افزارهای مشابه رایج، مشخص شد که سامانه پارسی اسپل به‌عنوان ابزار مؤثری جهت شناسایی و پیشنهاد واژه‌های صحیح برای خطاهای غیرواژه و واژه حقیقی است. در مراحل شناسایی و پیشنهاد، معیار-F به‌صورت معناداری بهبود یافته است. همچنین نتایج ارزیابی نشان داده که سامانه پارسی اسپل خطاهای واژه حقیقی بیشتری را شناسایی کرده و قادر به ارائه و پیشنهاد واژه‌های جایگزین صحیح، برای واژه‌های نادرست است و مقدار معیار بازخوانی در شناسایی خطای واژه حقیقی به‌صورت معناداری بیشتر از نرم‌افزارهای رقیب آن است.

واژگان کلیدی: سیستم خطایاب فارسی، تصحیح خطای واژگان، شناسایی خطای واژگان، پردازش زبان طبیعی، مدل زبان فارسی

Design and implementation of Persian spelling detection and correction system based on Semantic

Mohammad Bagher Dastgheib*[1], Sara Koleini[2] & Seyed Mostafa Fakhrahmad[3]

[1]Faculty member of Regional Information Center for Science and Technology (RICEST)

[2]Master of Science in Computer Engineering, RICEST

[3]Faculty member of School of Electrical and Computer Engineering, Shiraz University

Abstract

Persian Language has a special feature (grapheme, homophone, and multi-shape clinging characters) in electronic devices. Furthermore, design and implementation of NLP tools for Persian are more challenging than other languages (e.g. English or German). Spelling tools are used widely for editing user texts like emails and text in editors. Also developing Persian tools will provide Persian programs to check spell and reduce errors in electronic texts. In this work, we review the spelling detection and correction methods, especially for the Persian language. The proposed algorithm consists of two steps. The first step is non-word error detection and correction by intelligent scoring algorithm. The second step is read-word error detection and correction. We propose a spelling system "Perspell" for Persian non-word and real-word errors using a hybrid scoring

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

● تاریخ ارسال مقاله: ۱۳۹۶/۲/۱۸ ● تاریخ پذیرش: ۱۳۹۸/۳/۲۹ ● تاریخ انتشار: ۱۳۹۸/۱۰/۰۷ ● نوع مطالعه: کاربردی

[DOI: 10.29252/jsdp.16.3.128]

system and optimized language model by lexicon. This scoring system uses a combination of lexical and semantic features optimized by learning dataset. The weight of these features in scoring system is also optimized by learning phase. Perspell is compared with known Persian spellchecker systems and could overcome them in precision of detection and correction. Accordingly, the proposed Persian spell-checker system can also detect and correct real-word errors. This open challenge category of spelling is a complicated and time consuming task in Persian as well as, assessing the proposed method, the F-measure metric has improved significantly (about 10%) for detecting and correcting Persian words. In the proposed method, we used Persian language model with bootstrapping and smoothing to overcome data sparseness and lack of data. The bootstrapping is developed using a Persian dictionary and further we used word sense disambiguation to select the correct related replaced word.

Keywords: Spell Error Detection, Spell Error Correction, Persian spell Checker, NLP, Persian Language Model

۱- مقدمه

کاربرد گسترده رایانه‌ها در زندگی بشر باعث شده است که همواره مسائل و چالش‌هایی در پردازش زبان طبیعی مد نظر باشد. سامانه‌های تصحیح املای خودکار در حوزه‌های مختلفی از قبیل تصحیح پرس‌وجوهای یک عبارت پرس‌وجو، بررسی املایی واژگان در مرورگرها و برنامه‌های ویراستاری متنی کاربرد دارد. تصحیح خودکار واژه در متن از دهه ۱۹۶۰ تاکنون به‌عنوان چالشی در پژوهش‌های پردازش زبان طبیعی مطرح بوده و در همین راستا فنون و الگوریتم‌های بسیاری در حوزه‌های تجاری و علمی ایجاد شده است. هم‌اینک استفاده از روش های تصحیح املایی متون با رشد رسانه‌های اجتماعی در شبکه اینترنت بیش‌ازپیش مطرح است. درواقع بستر متنی وب از قبیل وبلاگ‌ها ،میکروبلاگ‌ها، انجمن‌ها، ... به‌عنوان منابعی جهت انجام پژوهش در این حوزه معرفی می‌شوند. بیش‌تر این بسترها بسیار بزرگ بوده، بنابراین اغلب به‌صورت خودکار جمع‌آوری و پردازش شده و تنها تصحیح دستی به‌صورت محدودی بر روی آنها انجام می‌پذیرد؛ ازاین‌رو به‌طورمعمول املای واژگانی که در متون این قبیل بسترها وجود دارد، دارای املاهای گوناگونی هستند؛ بنابراین توجه ویژه پژوهش‌گران به موضوع تصحیح املایی خودکار متون جلب شده است [1]. استفاده مؤثر از راه‌کارهای بررسی و تصحیح املای متون در سامانه‌های پردازش واژگان، سامانه‌های بازیابی اطلاعات، برنامه‌های تصحیح دستور زبان، برنامه‌های بازشناسی نوری نویسه‌ها (OCR) به اثبات رسیده [2] که با افزایش مهارت در برنامه‌های تصحیح خطاهای املایی واژگان، کارایی این گونه از برنامه‌ها بهبود می‌یابد. بیش‌تر پژوهش‌گران خطاهای املایی را به دو گروه عمده زیر تقسیم می‌کنند:

۱. خطاهای غیر واژه[1]- شامل واژگانی هستند که فاقد معنی بوده و در واژه‌نامه‌ها وجود ندارند.

۲. خطاهای واژه حقیقی[2]- در این نوع خطا، واژه بررسی شده در واژه‌نامه وجود دارد اما مورد نظر نویسنده نبوده و از این‌رو باعث می‌شود که جمله از لحاظ دستوری یا معنایی دارای خطا شود.

(جدول-۱): مثالی از انواع خطاهای مفرد در زبان فارسی
(Table-1): Example of Single-Error types

انواع خطاهای مفرد	مثال- واژه "سیب"
حذف	سب
درج	سیبج
تعویض	شیب
جابه‌جایی	سبی

در هر دو گروه باید روش شناسایی خطای واژگان (خطایابی واژگان) یا پیشنهاد جایگزینی واژه صحیح ارائه شود [3]. فنون شناسایی خطاهای غیرواژه به دو گروه عمده تحلیل مدل‌های چندوزنی و استفاده از واژه‌نامه‌ها تقسیم‌بندی می‌شود [2].خطاهای غیر واژه به دو گروه اصلی خطاهای مفرد[3] و خطاهای چندگانه[4] تقسیم‌بندی شده که در خطاهای مفرد تنها یک حرف که در واژه اضافه،کاسته، تعویض یا جابه‌جا شده باعث بروز خطا در واژه می‌شود. این نوع خطا با ذکر مثالی در جدول (۱) نشان داده شده است. میتون معتقد است که بیش از هشتاددرصد خطاها در واژه‌هایی با یک خطای ویرایشی از این گروه رخ می‌دهد [4]. در خطاهای چندگانه، واژهٔ دارای خطا، واژه‌ای است که دارای اشتباه نوشتاری بوده و در آن چند نمونه خطا رخ داده است [5]. روش‌هایی را که جهت خطایاب زبان تاکنون به‌کارگرفته شده است، می‌توان به دو گروه اصلی "روش‌های مبتنی بر قاعده" برای شناسایی

[1] Non-word
[2] Real-word
[3] Single errors
[4] Multi-error

[DOI: 10.29252/jsdp.16.3.128]

خطاهای دستوری [6] و "روش‌های آماری" برای شناسایی خطاهای املایی تقسیم‌بندی کرد [7].

۱-۱- راه‌کارهای شناسایی خطا

در طول چند دههٔ گذشته، فنون بسیاری جهت شناسایی و تصحیح خطاهای املایی پیشنهاد و برخی از آنها عملیاتی شده است. از راه‌کار اصلی که جهت شناسایی خطا استفاده می‌شود، می‌توان به فنون جستجو در واژه‌نامه‌ها و تحلیل مدل‌های چندوزنی[1]، اشاره کرد .دراین فنون اغلب به بررسی واژه‌به‌واژهٔ واژگان به‌منظور مشاهدهٔ خطاها پرداخته می‌شود [8].در ادامه شرح مختصری از روش‌های موجود جهت شناسایی خطا ارئه می‌شود:

- **تحلیل مدل‌های چندوزنی**

در این روش به‌جای مقایسه واژه موجود در متن با واژه‌نامه، فقط مدل‌های چندوزنی بررسی و کنترل می‌شود. این بررسی با ماتریس nبُعدی که در آن فراوانی‌های چندوزنی واقعی از متون تولید شده، توسط کاربران استخراج و ذخیره شده است، انجام می‌پذیرد. اگر یک چندوزنی نادر یا غیرموجود پیدا شد، در آن‌صورت واژه به‌عنوان واژهٔ دارای خطای املایی شناسایی می‌شود. یک چندوزنی، مجموعه‌ای از نویسه‌های متوالی برگرفته از یک رشته بوده که در این عبارت رشته‌ای، n برابر با طول رشته است. یکی از مزایای کاربرد الگوریتم چندوزنی در شناسایی خطا، مجازبودن رشته‌هایی با پیشوندهای متفاوت جهت تطابق و همچنین تحمل‌پذیری الگوریتم در برابر خطاهای املایی است [8]. مزیت عمده الگوریتم چندوزنی آن است که نیاز به دانشی از زبانی که استفاده شده است، ندارد؛ ازاین‌رو ازآن به‌عنوان الگوریتم مستقل از زبان یا الگوریتم تطبیق رشته خنثی یاد می‌شود. برای مثال استفاده از چندوزنی جهت محاسبه شباهت بین دو رشته استفاده می‌شود. بدین‌ترتیب که چندوزنی‌های واحدی که در هر دو رشته مشترک است، شناسایی شده و سپس ضریب تشابه محاسبه می‌شود. ضریب تشابه با تقسیم تعداد چندوزنی‌های مشترک، تقسیم بر تعداد کل چندوزنی‌ها در دو رشته به‌دست می‌آید.

- **استفاده از واژه‌نامه‌ها**

به‌منظور استفاده از واژه‌نامه‌ها در شناسایی خطا لازم است که از واژه‌نامه‌ای با اندازهٔ مناسب (دارای تعداد کافی از واژگان جهت مقایسه متون) استفاده شود. واژه‌نامه خیلی کوچک می‌تواند کاربر را با واژه‌های بسیاری که به‌اشتباه به‌دلیل عدم وجود در واژه‌نامه، خطا تشخیص داده شده است، روبه‌رو کند. به این واژه‌ها، واژه‌های خارج از واژه‌نامه گفته می‌شود؛ همچنین یک واژه‌نامه بسیار بزرگ نیز می‌تواند کاربر را با واژه‌های صحیح دور از ذهن که در زبان رایج نیست مواجه کند. بنابراین اندازه واژه‌نامه مهم است [9]. برای جستجوی سریع در واژه‌نامه‌ها از جداول درهم‌سازی[2] جهت دسترسی سریع به واژه استفاده می‌شود. به‌منظور جستجوی یک رشته، ابتدا نشانی درهم‌سازی آن محاسبه و سپس واژه‌ای که در آن نشانی ذخیره شده بازیابی می‌شود. چنانچه واژهٔ ذخیره‌شده در نشانی درهم‌سازی با رشتهٔ ورودی مورد نظر متفاوت بود در آن‌صورت خطای املایی صورت گرفته است. مزیت عمده جداول درهم‌سازی، ویژگی دسترسی تصادفی آن است که مقایسه‌های زیاد مورد نیاز جهت جستجو در واژه‌نامه را حذف می‌کند. عیب عمده آن، نیاز به وسیله‌ای جهت انجام تابع درهم‌ساز هوشمند جهت جلوگیری از برخورد است. جهت ذخیرهٔ یک واژه در واژه‌نامه، تابع درهم‌ساز واژه محاسبه و مقدار بردارهای متناظر برابر با مقدار "صحیح" در نظر گرفته می‌شود. جهت جستجوی یک واژه در واژه‌نامه، مقادیر درهم‌ساز واژه، محاسبه و با مقدار بردار مقایسه می‌شود [8].

۲-۱- راه‌کارهای تصحیح خطا

تصحیح خطا از موضوع‌های پژوهشی است که سال‌ها مورد توجه پژوهش‌گران است. بیشتر راه‌کارهای موجود، معنای واژه را در جمله در نظر نمی‌گیرد و فقط به املای واژه، جهت تصحیح واژه توجه می‌شود. فرایند اصلاح واژه شامل دو مرحله است. ابتدا تولید واژهٔ نامزد تصحیح‌شده و سپس رتبه‌بندی نامزدهای به‌دست‌آمده. در ادامه به معرفی چند راه‌کار اصلی می‌پردازیم:

- **روش مبتنی بر چندوزنی:**

این روش به دو روش استفاده از واژه‌نامه و یا بدون استفاده از واژه‌نامه تقسیم می‌شود. در روش استفاده از واژه‌نامه، از چندوزنی جهت مشخص‌کردن فاصله میان واژه‌ها استفاده شده، اما بررسی واژه با واژه‌نامه انجام می‌گیرد. در روشی که از واژه‌نامه استفاده نمی‌شود، محل رخداد خطا در واژه تعیین می‌شود. درصورتی‌که واژه‌ای را که دارای خطای املایی است، بتوان به‌گونه‌ای تغییر داد که فقط شامل چند وزنی‌های معتبر باشد، در آن صورت تصحیح واژه انجام گرفته است. مزیت عمده این روش سادگی و عدم نیاز به واژه‌نامه است [9].

[1] N-Gram

[2] Hash Table

• **روش‌های مبتنی بر قاعده:**

در این روش مجموعه‌ای از قوانین که شامل خطاهای املایی رایج و خطاهای چاپی است به واژه دارای خطا اعمال می‌شود. این قوانین دارای احتمالاتی بوده که باعث رتبه‌بندی پیشنهادها با جمع‌آوری احتمالات برای قوانین اعمال‌شده می‌شود [10].

• **شبکه‌های عصبی:**

یکی از روش‌ها، استفاده از شبکه عصبی پس‌انتشار خطا است. در این روش یک گره ورودی برای هریک از چند وزنی‌ها که در موقعیت‌های مختلف واژه قرار دارند، در نظر گرفته شده و از یک گره خروجی برای هر واژه موجود در واژه‌نامه استفاده می‌کند. به‌طورمعمول تعداد لایه‌های پنهان برابر "یک" یا"دو" در نظر گرفته می‌شود [8].

• **روش کمترین فاصلهٔ ویرایشی:**

این روش، بیش‌ترین کاربرد را در پژوهش‌ها داشته است. این دسته از الگوریتم‌ها، کمینه فاصلهٔ بین رشتهٔ دارای خطا املایی و رشته‌های درون واژه‌نامه را محاسبه می‌کنند. نام کمترین فاصلهٔ ویرایشی مطرح شده است [5,11] و به‌صورت "تعداد کمترین عمل‌گرهایی (درج-حذف-جایگزینی-جابه‌جایی) که لازم است یک رشتهٔ خطا را به رشتهٔ درست تبدیل کند." تعریف می‌شود. این روش ساده بیشتر برای تصحیح خطاهای ورودی واژگان از طریق صفحه‌کلید به‌دست می‌آید. این روش برای اصلاح خطاهای املایی آوایی مناسب نیست به‌خصوص اگر تفاوت، بین املا و تلفظ باشد [5]. روش دیگر فاصله ویرایشی کاشفی است [3] که در این روش، محل قرارگیری حروف در چیدمان صفحه‌کلید، هم‌آواییِ حروف، هم‌شکل‌بودن حروف، تأثیر کلید شیفت، تنوین و حرکت‌ها را نیز در فاصلهٔ رشته‌ای در نظر گرفته می‌گیرد.

• **روش‌های احتمالی:**

این روش بر اساس برخی از ویژگی‌های آماری زبان است که به دو روش احتمال جابه‌جایی[1] (که مشابه چندوزنی بوده) و احتمال آشفتگی[2] تقسیم می‌شود. روش احتمال جابه‌جایی در زمانی که به واژه‌نامه دسترسی وجود دارد مناسب نیست. جهت تصحیح یک جمله، سامانه جمله را به چند واژه تجزیه کرده و نامزدهای مناسب واژه را از واژه‌نامه بازیابی و با رشته چندوزنی مقایسه می‌کند. نامزدهای بازیابی‌شده به‌وسیله احتمال شرطی تطابق با رشته رتبه‌بندی‌شده و احتمال آشفتگی نویسه را رقم می‌زند؛ سرانجام یک مدل دووزنی واژه و یک الگوریتم ویژه جهت تعیین بهترین نمره ترتیب واژگان در جمله به‌کار گرفته می شود. ادعا می‌شود که این روش می‌تواند خطاهای غیرواژه‌ای را به‌خوبی خطاهای واژه حقیقی تصحیح کرده و به نرخ کاهش خطای ۶۰/۲٪ برای متن شناسه نوری نویسه‌های حقیقی دسترسی یابد [8].

۲- پیشینهٔ پژوهش

مطالعه‌ای جهت ارزیابی اثربخشی آمار فراوانی سه‌وزنی برای شناسایی خطاهای املایی، بررسی واژه‌ها با املای صحیح و ایجاد تمایز میان انواع اصلی خطاهای املایی بر روی بستری شامل پنجاه‌هزار واژهٔ گردآوری‌شده از هفت پایگاه اطلاعاتی در حوزهٔ شیمی انجام شده است [12]. این پژوهش‌گران دریافتند که روش تحلیل سه‌وزنی قادر به مکان‌یابی خطای یک حرف در واژه‌ای که دارای خطای املایی است در ۹۴٪ موارد است. مدل تلفظ برای تصحیح املا [13]، و هم‌چنین مدلی شامل ویژگی‌های آوایی درخصوص بررسی و تصحیح خطای املایی متون ارائه شده است [14]. در پژوهشی دیگر مشاهده شد که خطای پیوستگی و انفصال در اغلب موارد می‌تواند منجر به تولید یک واژهٔ معتبر در واژه‌نامه شود [15]. برای مثال واژهٔ inform می‌تواند به دو واژهٔ معتبر "in" و "form" تبدیل شده و یا "سلامتی" می تواند به یک واژهٔ معتبر "سلام" و یک واژهٔ نامعتبر "تی" تبدیل شود. خطای پیوستگی و انفصال نشان می‌دهد که تشخیص محدودهٔ واژه‌ها بسیار مهم است. به‌طورمعمول خطایاب‌ها تشخیص محدودهٔ واژه‌ها را به‌عنوان یک خطای جداگانه در نظر نمی‌گیرند. به‌طورکلی، در زمان تصحیح خطا در صورت کمتربودن مقدار پیشنهاد از مقدار آستانه، خطای پیوستگی و انفصال بررسی می شود[15]. در پژوهش دیگری، فرایند تصحیح خطا را به چهار گروه عمده نشان‌گذاری، شناسایی خطا، تصحیح خطا و مرتب‌سازی و ارائه پیشنهادها تقسیم شده که تصحیح خطا یکی از برنامه‌های کاربردی در پیکره‌های بزرگ تک‌زبانی است [16]. از این‌رو وجود پیکره‌های بزرگ و ابزار پردازش زبان طبیعی از ملزومات پژوهش در این حوزه محسوب می شود. پژوهش‌هایی در مورد استفاده از روش‌های یادگیری ماشین در مدل‌سازی سامانه بررسی خطا صورت گرفته است [17,18]. چند ابزار پردازش زبان طبیعی برای متون فارسی ارائه داده شده است؛ از جمله ویراستیار، که نرم‌افزار خطایاب املایی فارسی بوده

[1] Transition Probability
[2] Confusion Probability

[DOI: 10.29252/jsdp.16.3.128]

که قادر به شناسایی خطاهای غیرواژه‌ای مختلف است [19,20]. این پژوهش‌گران، گزارش دادند که واژه‌های بررسی شده به‌تنهایی چندان مؤثر نبوده و مسایل مشترکی در اغلب خطایاب‌های املایی فارسی وجود دارد. این مسایل دارای دو راه حل متفاوت - یعنی بررسی قوانین دستورزبان فارسی و استفاده از مدل آماری به منظور بررسی اصطلاحات - است. راه حل نخست پرهزینه و پیچیده بوده، زیرا دستور زبان فارسی پیچیده است [20]. سامانه خطایاب املایی فارسی دیگری به نام وارسی‌گر فارسی (وفا) وجود دارد که خطاهای غیرواژه‌ای، خطاهای دستورزبانی و خطاهای واژه حقیقی را شناسایی می‌کند [21]. خطایاب املایی وفا بر اساس یک واژه‌نامه حجیم که شامل واژه‌های رایج زبان فارسی است، انجام می‌شود. در این واژه‌نامه تمام واژگان زبان، ریشه‌ها و مشتقات آنها که از نظر زبان، مرسوم و صحیح‌اند، موجود است. تصمیم‌گیری در مورد تشخیص واژگان نادرست (دارای خطاهای نوشتاری) نیز بر اساس این واژه‌نامه انجام می‌گیرد. بدین ترتیب، واژه‌ای که در واژه‌نامه موجود نباشد، به‌احتمال دارای خطای نوشتاری بوده و باید برای اصلاح آن اقدام کرد. خطاهای دستوری به‌صورت مبتنی بر قانون و با تعریف یک سری قواعد دستور زبانی در سطح جملات کنترل می‌شوند. به‌منظور تشخیص و تصحیح خطاهای معنایی نیز الگوریتم‌ها و روش‌های متعددی به‌کار گرفته شده است. نتایج با استفاده از توزیع احتمال چند وزنی بر واژه ها بهبود می‌یابد.

۳- ویژگی‌ها و چالش‌های زبان فارسی از دیدگاه پردازش الکترونیکی

زبان فارسی زبان رسمی ایران، تاجیکستان و یکی از دو زبان عمده‌ای است که در افغانستان استفاده می‌شود. زبان فارسی وام‌دار واژه‌های زیادی از زبان‌های دیگر مانند عربی است؛ اما در طی سال‌ها، ساختار زبان فارسی حفظ شده و تغییری ساختاری در زبان فارسی صورت نپذیرفته است [20]. زبان فارسی دارای ساختار واژگانی آزاد در جملات است، برای مثال قید می‌تواند در ابتدای جمله، در میان جمله یا در آخر جمله قرار گیرد. در زبان فارسی، حرف نخست جمله مانند زبان انگلیسی، حرف بزرگ نیست، بنابراین تشخیص ابتدا و انتهای جملات، خود چالش جدیدی در زبان فارسی است [22]. به‌عبارت دیگر شکستن قطعه‌ای از متن به پاراگراف و جمله در زبان فارسی بسیار چالش‌برانگیز است و ابزارهای استانداردی برای آن وجود ندارد. ساختار نویسه‌های فارسی، راست به چپ است و از الفبای عربی به‌علاوه چهار حرف اضافهٔ (پ،ژ،گ،چ) برای نوشتار بهره می‌برد. زبان فارسی هم‌چنین دارای مصوّت‌های کوتاه است، که اغلب در نوشتار حذف شده و نوشته نمی‌شوند. حذف این مصوّت‌های کوتاه، ممکن است، مشکلاتی را برای تلفظ واژه‌ها، و یا حتی ابهام در معانی واژه‌ها ایجاد کند. براساس اطلاعات به‌دست‌آمده از ویکی‌پدیا، ۱٪ جمعیت جهان فارسی را به‌عنوان زبان مادری صحبت می‌کنند و در حدود ۱۳۴ میلیون نفر از مردم جهان، فارسی را به‌عنوان زبان نخست یا دوم صحبت می‌کنند. مخاطبان گستردهٔ زبان فارسی در سطح جهان نیازمند انجام پژوهش در حیطهٔ پردازش زبان فارسی هستند. از دیگر چالش‌های زبان فارسی، واژه‌شناسی غنی و پیچیده آن است [18]. واژه‌های زبان فارسی با ترکیب‌های بسیاری از پسوندها و پیشوندها صرف می‌شوند. واژگان اشتقاقی بسیاری در زبان فارسی رایج است و به کار برده می‌شود، ولی قوانین اشتقاق، تصریف، و ترکیب آنها دقیق و جامع نیست. ترکیب وندها با اسامی در زبان فارسی یکی از مشکلات شناخت واژه‌ها است؛ ولی به‌طور بایسته به آن پرداخته نشده و قوانین جامعی برای آن تدوین نشده است [20]. عدم وجود قوانین، موجب برخوردهای متفاوت و سلیقه‌ای شده و پژوهش‌گران حوزهٔ پردازش زبان طبیعی را سردرگم می‌کند. افعال فارسی نیز از نبود قوانین تعریف‌شدهٔ مشخص برای صرف و ترکیب افعال رنج می‌برند؛ بنابراین وجود افعال مرکب، نبود قوانین مدون، فعل‌های پیچیدهٔ چندجزئی و همچنین وجود استثناهای فراوان از جمله چالش‌های افعال فارسی است [20]. زبان فارسی علاوه‌بر فاصله‌گذاری معمول در دیگر زبان‌ها که به‌عنوان جداساز[1] واژگان استفاده می‌شود، فاصلهٔ دیگری به نام فاصلهٔ درون‌کلمه‌ای و یا شبه‌فاصله[2] دارد. شبه‌فاصله‌ها دارای قوانین مدون و دقیقی نیستند [22]. با توجه به تمامی موارد بالا، تشخیص ترتیب و محدوده واژه‌ها در یک جمله خود یکی از مهمترین چالش‌های زبان فارسی است [19] . برای زبان فارسی ابزار دقیق و مستندی برای تجزیهٔ جمله به واژه وجود ندارد و یا به‌طور عمومی در دسترس پژوهش‌گران نیست و یا در صورت انتشار مشکلات زیادی دارد. این ابزار، مورد نیاز بسیاری از پژوهش‌ها در حیطهٔ پردازش‌های زبان طبیعی است و یک پیش‌نیاز پژوهش در حوزهٔ پردازش زبان طبیعی محسوب می‌شود.

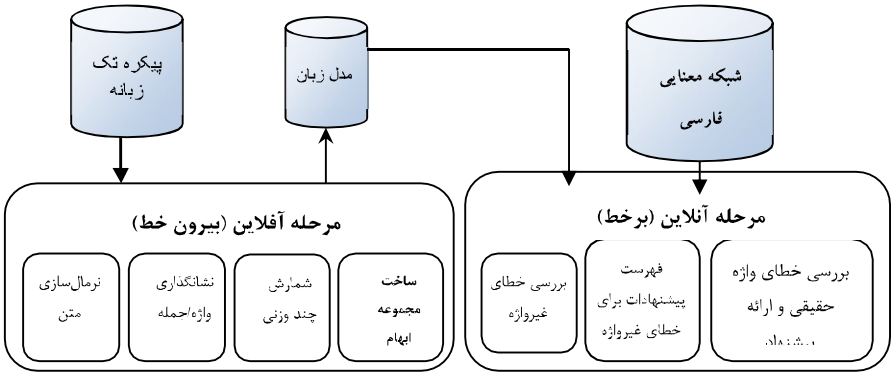
[1] Delimiter

[2] Pseudo-space or half-space (نیم فاصله)

[DOI: 10.29252/jsdp.16.3.128]

۴- سامانهٔ خطایاب املایی فارسی پارسی‌اسپل

در این پژوهش چالش‌های زبان فارسی از دیدگاه نوشتاری املای واژه‌ها، و همچنین پردازش رایانه‌ای مورد بررسی قرار گرفته و راه‌کارهای آماری در پیکره تک‌زبانی اعمال شده است. به‌منظور شناسایی خطاهای واژه حقیقی، از متنی که شامل واژه موردنظر (هدف) است، استفاده شد. یکی از مهم‌ترین دستاوردهای این پژوهش، معرفی مدلی برای زبان فارسی است که احتمال ترکیب‌های واژه را در یک متن مشخص می‌کند. با درنظرگرفتن ویژگی‌های زبان فارسی، سامانه خطایاب املایی پارسی‌اسپل طراحی شد. شکل (۱) مرور مختصری را بر معماری سامانه ارائه می‌دهد. وظایف این سامانه به دو گروه اصلی یعنی برخط و برون‌خط تقسیم می‌شود. در قسمت برون‌خط، تمام داده‌هایی که باید در فرایند بررسی خطای املایی استفاده شود، تولید می‌شود. در این قسمت نرمال‌سازی انجام شده و مسائلی که در پردازش زبان فارسی وجود دارد، مانند شبه‌فاصله‌ها، فضاهای خالی و پسوندها را حذف می‌کند [16]. برخی از فعالیت‌های پیچیده و پرهزینه نظیرساخت مجموعه‌های ابهام و مدل چندوزنی زبانی در این قسمت انجام می‌شود. در قسمت برخط فرایند بررسی املا به دو قسمت عمده تقسیم می‌شود. در بخش نخست، شناسایی خطاهای غیرواژه‌ای و ایجاد فهرست‌های پیشنهادی انجام گرفته و در قسمت دوم خطاهای واژه حقیقی شناسایی شده و پیشنهادها در مجموعه ابهام تنظیم می‌شود. هر دو مرحله از دانش و داده‌های تولیدشده در قسمت برون‌خط استفاده می‌کنند. در ادامه، نحوه ساخت و کارکرد هر یک از ماجول‌های اصلی این سامانه شرح داده می‌شود.



(شکل-۱): معماری پارسی اسپل

(Figure-1): The Architecture of پارسی اسپل

۴-۱- ساخت واژه‌نامه از بستر تک‌زبانه

مهم‌ترین مرحله هر سامانه خطایاب املایی فارسی، جمع‌آوری تمام اشکال صحیح واژگان در یک فرمت دیجیتالی است. در این پژوهش از متن چکیده بیش از چهارصدهزار مقاله علمی-پژوهشی وزارت علوم تحقیقات و فناوری و وزارت بهداشت برای تولید پیکره به‌عنوان منبع استفاده شده است. از آنجا که متون مقالات و به‌خصوص چکیده‌ها، ممیزی می‌شوند، بنابراین خطای واژه‌ای و دستوری در آن‌ها با نرخ بسیار کمتری نسبت به متون روزنامه و یا متون موجود در وب یافت می‌شود؛ ولی چالش‌های موجود درخط فارسی موجب بروز برخی خطاها در متن شده، لذا این متون نیاز به پیش‌پردازش دارند. انتخاب چکیده مقالات با توزیع نرمال به‌صورت تصادفی از تمامی موضوع‌های در دسترس (فنی و مهندسی، علوم پایه، علوم انسانی و پزشکی) از مخزنی با بیش از هشتصدهزار مقاله انتخاب شده است. درضمن متن مقالات، متن رسمی است، ولی متن به‌کاررفته در تارنماها و روزنامه‌ها اغلب متن غیر رسمی و یا عامیانه بوده که صورت رسمی و صحیح واژه‌ها را به‌کار نمی‌گیرد و درنتیجه پیکرهٔ به‌دست‌آمده سطح نوفهٔ بالایی دارد. (منظور از نوفه، خطاهایی است که در پیکره برای واژگان وجود دارد). در مرحله بعد، مقالات به متن ساده[1] تبدیل شده و همه جملات نشانه‌گذاری می‌شوند. واژه‌های رایج با واژه‌نامه دهخدا بررسی و از واژگانی که دارای فراوانی کمتر از آستانه مورد نظر (فرکانس تکرار ۱۰) هستند، صرف‌نظر می‌شود؛ سپس به‌صورت دستی واژه‌های خارج از واژه‌نامه مورد بررسی قرار می‌گیرند. با گذر زمان واژه‌های موجود در هر زبانی تغییر کرده و واژگانی از طریق زبان‌های خارجی به آن افزوده می‌شود. از این‌رو روزآمدسازی واژه‌نامه

[1] Plain Text

[DOI: 10.29252/jsdp.16.3.128]

حائز اهمیت است. واژه‌نامه نهایی به‌دست‌آمده از این پژوهش شامل یک‌میلیون و چهارصد هزار واژه فارسی به‌همراه واژه‌های اضافه‌شده از سایر زبان‌ها است. جهت تصحیح و پیشنهاد جایگزین‌ها برای خطاهای غیرواژه‌ای از فاصله دمراو- لونشتاین استفاده شد. جهت ساخت مجموعه ابهام، تمام عملیات ویرایشی از جمله حذف، درج، جابه‌جایی، تعویض، و برخی از ویژگی‌های زبان فارسی مانند هم‌آوایی[1] در نظر گرفته می‌شود. همچنین از فاصله ویرایشی کاشفی [3] که مجاورت کلیدهای صفحه کلید فارسی را در نظر می‌گیرد، استفاده شد؛ علاوه‌براین، فضای سفید را به‌عنوان یک نویسه گم‌شده و یا حذف‌شده در نظرگرفته تا واژگانی را که به‌اشتباه متصل شده، تشخیص دهد. سرانجام، مجموعه ابهام به کاربر نشان داده می‌شود تا واژه‌هایی را با املای نادرست، جایگزین کند. پیچیدگی زمانی جستجو در واژه‌نامه قابل قبول است؛ زیرا جستجو در ساختار درهم‌سازی واژه‌نامه با پیچیدگی زمانی O(1) انجام شده و رتبه پیچیدگی بررسی یک جمله که شامل M واژه است، برابر با O(M) است.

۲-۴- شناسایی خطای واژه حقیقی و ساخت مجموعه ابهام

خطای واژه حقیقی به‌طورمعمول از طریق واژه‌پرداز، ضمن فرایند "تصحیح خودکار" به‌وجود می‌آید [23]. ممکن است این خطاها توسط انسان و ماشین تولید شود. در این پژوهش، واژه‌ای از مجموعه ابهام که دارای احتمال کمتری است، به‌عنوان خطای واژه حقیقی در نظر گرفته می‌شود [23]. روش پیشنهادی برای هر اصطلاح نامزدشده، مجموعه ابهامی از واژه‌ها را تشکیل می‌دهد که همگی دارای فاصله دمراو- لونشتاین برابر با یک برای هر واژه موردنظر (هدف) است. به‌عبارت دیگر هر واژه به‌طورمعمول دارای رخداد بیش از یک خطا نیست. مجموعه ابهام به‌وسیله افزودن مترادف‌های هر واژه که توسط روش ابهام‌زدایی حساس به واژه[2] در موضوع هر جمله پالایش شده است، بسط می‌یابد؛ سپس واژگان موجود در مجموعه ابهام مرتب شده و با ملاحظه حد آستانه (۱۰ واژه برتر) پالایش می‌شود. درصورتی‌که واژه موردنظر (هدف) در فهرست نتیجه موجود نباشد، آنگاه به‌عنوان خطای واژه حقیقی علامت‌گذاری می‌شود. فهرست رتبه‌بندی‌شده به کاربر به‌عنوان فهرست پیشنهادی جهت تصحیح خطا نشان داده می‌شود.

[1] منظور آوای یکسان نویسه‌های مختلف نظیر Homophones ذ،ز،ظ،ض است.

[2] Word- sense disambiguation(WSD)

۱-۲-۴- مدل چندوزنی واژگانی فارسی

جهت مرتب‌سازی واژگان مجموعه ابهام، از راه‌کارهای نمره‌دهی که براساس احتمالات مدل زبانی است، استفاده می‌شود. به‌منظور تشکیل مدل زبانی، فراوانی‌های دووزنی از متن چکیده مقاله استخراج می‌شود. برای هر واژه W^i موجود در متن، اگر W^{i1} واژه قبلی و W^{i-1} واژه بعدی باشد، در آن‌صورت دووزنی چپ و راست واژه به‌ترتیب به‌صورت $W^{i1}W^i$ و W^iW^{i+1} تولید می‌شود. در مدل چندوزنی، احتمال تخصیص داده‌شده به دنباله‌ای که شامل n واژه است، به‌صورت معادله (۱) محاسبه می‌گردد:

$$P(w^1, w^2, \dots, w^n) = \prod_{i=1}^{n} P(w^i | w^1 \dots w^{i-1}) \qquad (۱)$$

همچنین می‌توان مدل زبانی را به‌وسیله مدل زنجیره مارکوف[3] تخمین زد. در قانون مارکوف فرض می‌شود که احتمال یک رویداد (برای مثال واژه بعدی در متن) تنها به تاریخچه محدود رویدادهای پیش از آن (برای مثال، دو یا سه واژه قبلی) بستگی دارد [23]. در معادله (۲)، از مدل زبانی دووزنی جهت محاسبه احتمال جمله استفاده شده است.

$$P(w^1, w^2, \dots, w^n) = \qquad (۲)$$
$$P(w^1 | < S >) P(w^2 | w^1)$$
$$\dots P(w^n | w^{n-1}) P(</S > | w^n)$$

در معادله (۲) فرض می‌شودکه رویداد یک واژه نمونه تنها به واژگان بعدی و قبلی آن بستگی دارد. از معادله (۳) برای محاسبه و تخصیص نمرات به تمام اعضای مجموعه ابهام استفاده می‌شود؛ سپس فرض می‌شود که واژه موردنظر (هدف)، خطای واژه حقیقی است که به‌وسیله هر عضو مجموعه ابهام جایگزین شده است. در هر حالت احتمال جمله محاسبه شده و سرانجام اعضای مجموعه ابهام بر اساس احتمالات محاسبه‌شده مرتب می‌شوند. جهت جلوگیری از مشکل پاریز[4]، قانون زنجیره‌ای را تغییر داده و به‌صورت لگاریتم احتمالات به‌صورت زیر بیان می‌شود:

$$Score(w_k^i) = \log\left(P(w^1 | < S >) P(w^2 | w^1) \dots P(w^n | w^{n-1}) P(</S > | w^n)\right) \qquad (۳)$$

[3] Markov chain rule
[4] Underflow

[DOI: 10.29252/jsdp.16.3.128]

$$Score(w_k^i) = \log(P(w^1| < S >)) + \log(P(w^2|w^1)) + \cdots + \log(P(w^n|w^{n-1})) + \log(P(</S > |w^n)) \quad (۴)$$

مدل الگوریتم پیشنهادی جهت مشخص‌کردن خطای واژه حقیقی‌بودن واژهٔ نمونه به‌کار برده می‌شود (شکل-۲). در الگوریتم خطایاب املایی واژه حقیقی_مقدار آستانه بر اساس مقدار بالاترین عضو مرتب‌شده از مجموعه ابهام در نظر گرفته می‌شود [23]. مقدار آستانه می‌تواند از حاصل‌ضرب عددی کمتر از یک در بالاترین نمره رتبه‌بندی‌شده موجود در مجموعه ابهام محاسبه شود. جدول (۲) مشخصه مدل زبان فارسی را در این پژوهش نشان می‌دهد. جمله "ویژگی شهر تهران آلودگی هوا است" را در نظر بگیرید. اگر واژه "شهر" به‌صورت خطا "شهد" نوشته شود، جملهٔ بی‌معنی خواهیم داشت. جابه‌جایی حرف "ر" با حرف "د" به‌عنوان خطای واژه حقیقی در نظر گرفته می‌شود. جدول (۳) اعضای مجموعه ابهام برای واژه "شهد" را نشان می دهد. مجموعه ابهام دارای ۲۷ عضو بوده که تنها چند عضو آن کمتر از حد آستانه هستند. جدول (۴) مجموعه نهایی پیشنهادی را نشان می‌دهد. فهرست پیشنهادی شامل ده واژه‌ای که دارای رتبه نخست در مجموعه ابهام بوده و کمتر از حدآستانه هستند، می‌شود. فهرست پیشنهادی به کاربر جهت انتخاب واژه موردنظر نشان داده می‌شود.

```
Algorithm RealWordErrorDetect
Input: W^i    // a typewritten word
Output: Boolean (True or False)//a real word error or not
{
  if( Score(W^i) = = 0)
= STEM(W^i);W^i_stem
      if(W^i_stem exists in lexicon)
          if(Score(W^i_stem) = = 0)
              return true;     // W^i is a real word error
          else return false;   // W^i is correct
      else return true;
  else
=STEM(W^i);W^i_stem
      if(W^i_stem exists in lexicon)
            if(Score(W^i_stem)< Threshold)
                return true;
            else  return false;
       else return true;
}
```

(شکل-۲): الگوریتم شناسایی خطای واژه حقیقی

(Figure-2): Real word error detection Algorithm

اگر واژه موردنظر در فهرست باشد، آن واژه به‌عنوان پیشنهاد موفق علامت‌گذاری می‌شود. چنانچه رتبه محاسبه‌شده برای واژه اصلی بالاتر از آستانه تعریف‌شده از قبل باشد، واژه به‌عنوان خطای واژه حقیقی علامت‌گذاری نمی‌شود. برخی از دنباله‌های نامزدشده که از مجموعه ابهام به‌دست آمده‌اند، ممکن است، دارای فراوانی دووزنی برابر با صفر در مدل زبانی شوند. جهت جلوگیری از وجود مقادیر صفر، از یک‌نواخت‌سازی و بسط واژه با درنظرگرفتن مترادف واژه‌ها استفاده می‌شود. از ابهام‌زدایی واژه‌ها جهت پالایش واژه‌های غیر مربوط بر اساس معنای محتوا استفاده می‌شود. نقش ابهام‌زدایی حساس به واژه‌ها در این پژوهش، کاهش نوفه واژه‌های غیر مربوط در مجموعه ابهام است. جهت بهبود الگوریتم براساس دقت[1] و بازخوانی[2]، هر دو واژه چندوزنی با و بدون ریشه‌سازی ثبت می‌شود.

(جدول-۲): ویژگی‌های مدل زبان فارسی استفاده‌شده در پارسی اسپل

(Table-2): Persian language model properties in پارسی اسپل

ویژگی	مقدار
تعداد چکیده‌ها	400,158
تعداد پاراگراف‌ها	1,400,533
تعداد جملات	7,703,041
تعداد نویسه‌ها	46,218,187
تعداد واژه‌های مجزا	1,044,568
تعداد دووزنی‌های مجزا	1,974,788

(جدول-۳): برخی از اعضای مجموعه ابهام

(Table-3): Some members of confusion set

واژه
شهد
شهر
عهد
مهد
شود

(جدول-۴): مجموعه نهایی پیشنهادی برای واژه "شهد"

(Table-4): Final suggestion for "شهد"

واژه	رتبه	نمره
شهر	1	-0.302
شاهد	2	-2.915

[1] Precision

[2] Recall

[DOI: 10.29252/jsdp.16.3.128]

۵- آزمون‌ها و نتایج

جهت ارزیابی مدل پیشنهادی، یکصد چکیدهٔ مقاله از مخزن استخراج شد. چکیدهٔ انتخاب‌شده از موضوعات مختلف با توزیع نرمال بود. مجموعه داده دارای ۲۷۰ پاراگراف، ۱۲۴۳ جمله، ۱۳۷۶۵ واژه و ۷۷۶۲۵ نویسه با احتساب فضای سفید است. فرایند ارزیابی در دو مرحله انجام گرفت. نخستین قسمت آزمون، عملکرد خطایاب املایی برای خطاهای غیرواژه بررسی می‌شود. به این ترتیب که برای هر چکیده موجود در مجموعه آزمایشی، چند واژه (۲۰٪ واژه‌ها) به‌صورت تصادفی انتخاب شده و با خطای غیر واژه جایگزین می‌شود. یکی از نکات مدنظردر این آزمون، شبیه‌سازی نحوه تایپ‌کردن کاربر حقیقی است تا مکان خطاها و واژه‌های اصلی جهت ارزیابی در نظر گرفته شود. عملکرد سامانه خطایاب املایی با استفاده از معیارهای دقت، بازخوانی و معیار-F ارزیابی می‌شود. مقادیر دقت و بازخوانی در ضمن مراحل شناسایی خطا و پیشنهاد واژه محاسبه می‌شود. نتایج شناسایی خطا در جدول (۵) ارائه و با خطایاب وفا [21] و ویراستیار [20] مقایسه شده است. جدول (۶) نتایج مرحله ارائه پیشنهاد را نشان می‌دهد. عملکرد سامانه‌های خطایاب املایی جهت شناسایی خطاهای غیرواژه به میزان زیادی به استفاده از واژه‌نامه بستگی دارد. این آزمون، با توجه به نتایج به‌دست‌آمده در جدول (۵)، نشان داده است که سامانه خطایاب املایی پارسی اسپل با توجه به مقادیر به‌دست‌آمده از معیارهای دقت و بازخوانی، از سایر سامانه‌ها از دقت بالاتری برخوردار است. گفتنی است که ساخت واژه‌نامه از مقاله‌های علمی فارسی در مقایسه با سامانه خطایاب وفا بسیار مؤثرتر و دقیق‌تر است. واژه‌نامه مورد استفاده در خطایاب وفا از روزنامه‌ها تولید شده و شامل واژه‌های خاص علمی نبوده و بنابراین دارای عملکرد ضعیفی در مقایسه با پارسی‌اسپل است. نتایج مرحله پیشنهادی جدول (۶) نشان می‌دهد که جهت انتخاب نامزد، فاصله ویرایشی یک مؤثر بوده و نیازی به درنظرگرفتن مقادیر بالاتر نیست؛ زیرا خطاهای غیرواژه با فاصله ویرایشی بالا به‌ندرت در متن مشاهده می‌شود. در قسمت بعدی آزمون، جهت ارزیابی عملکرد پارسی‌اسپل در شناسایی خطای واژه حقیقی، یکصد چکیده مقاله با موضوع‌های مختلف با توزیع نرمال از مخزن استخراج شد. جهت ساخت مجموعه آزمون برای خطای واژه حقیقی، به‌صورت تصادفی ۲۰٪ واژه‌ها در هر متن انتخاب شده و با واژهٔ جدید دارای دو محدودیت، جایگزین شد.

(جدول -۵):ارزیابی مرحله شناسایی خطای غیرواژه
(Table 5): Non-word error detection phase evaluation

	میانگین دقت	میانگین بازخوانی	معیار-F
پارسی اسپل	0.966164	1	0.98251
ویراستیار	0.965755	0.959147	0.961604
وفا	0.919883	0.944902	0.930643

(جدول- ۶): ارزیابی مرحله پیشنهاد خطای غیرواژه
(Table-6): Non-word error suggestion phase evaluation

	میانگین دقت	میانگین بازخوانی	معیار-F
پارسی اسپل	0.883061	0.888706	0.87748779
ویراستیار	0.835585	0.988056	0.723879593
وفا	0.801299	0.805312	0.797407

نخستین شرط آن است که واژه جدید باید در واژه‌نامه وجود داشته و واژه صحیحی باشد و شرط دوم دارابودن فاصله ویرایش "یک" از واژه مورد نظر (هدف) است. درصورتی‌که واژه‌ای با دو شرط بالا پیدا نشود، از واژه موردنظر (هدف) صرف‌نظر شده و واژه دیگری از متن انتخاب می‌شود. این فرایند تا زمانی‌که تعداد واژگانی که دارای خطای واژه حقیقی هستند به ۲۰٪ تمام واژگان موجود در متن برسد، ادامه می‌یابد. در این مرحله واژگان بازدارنده[1] فارسی و واژگانی که دارای طول کمتری از سه حرف باشد، در نظر گرفته نمی‌شود. مراحل ارزیابی این سامانه نیز مانند قبل است. ابتدا معیارهای دقت و بازخوانی برای مرحله شناسایی و سپس برای مرحله پیشنهادی محاسبه شد. نتایج به‌ترتیب در جدول‌های (۷ و ۸) نشان داده شده است. به نظر می‌رسد که تنها سامانه رایگان خطایاب املایی فارسی وفا می‌تواند خطاهای واژه حقیقی را شناسایی کند، از این‌رو نتایج حاصله از پارسی اسپل با نتایج به‌دست‌آمده از سامانه وفا مقایسه شد. نتایج نشان می‌دهد که عملکرد پارسی اسپل در محدوده متون علمی مطلوب است. اگرچه دقت پارسی اسپل کمتر از سامانه خطایاب وفا است، اما با مقایسه بازخوانی کل دو سامانه مشاهده می‌شود که تعداد خطاهای واژه حقیقی شناسایی شده در سامانه وفا بسیار کمتر از سامانه پارسی اسپل بوده که ۹۵٪ خطای واژه حقیقی را شناسایی می‌کند. برای هر مدرک در سامانه خطایاب وفا، میانگین تعداد خطاهایی که شناسایی نمی‌شود، برابر با ۲۳/۶ است. به طریق مشابه، برای پارسی اسپل تعداد خطاهایی که شناسایی نشدند تنها برابر با ۵/۱ است. با توجه به نتایج موجود

[1] Stop words

[DOI: 10.29252/jsdp.16.3.128]

در جدول (۷)، مشخص است که عملکرد پارسی‌اسپل در مقایسه با سامانه وفا به‌وسیله معیار F، در هر دو مرحله شناسایی و پیشنهاد به‌طور معناداری بهتر است.

(جدول-۷): نتایج ارزیابی مرحله شناسایی خطای واژه حقیقی
(Table-7): Real word error detection phase evaluation

	میانگین دقت	میانگین بازخوانی	معیار-F
پارسی اسپلسیستم	0.587288	0.969037	0.726091
سیستم وفا	0.815556	0.094179	0.165409

(جدول-۸): نتایج ارزیابی مرحله پیشنهاد خطای واژه حقیقی
(Table-8): Real word error suggestion phase evaluation

	میانگین دقت	میانگین بازخوانی	معیار-F
پارسی اسپلسیستم	0.923676	0.929545	0.926527
سیستم وفا	0.759259	0.986728	0.794433

۶- نتیجه‌گیری

در این مقاله، سامانه تصحیح خطایی که برای واژه‌های فارسی پیاده‌سازی شده است، معرفی شد. طرح پیشنهادی بر چالش‌های موجود در پردازش زبان فارسی تمرکز دارد. نتایج آزمون‌ها نشان می‌دهد که سامانه پارسی‌اسپل به‌عنوان ابزار مؤثری جهت شناسایی و پیشنهاد واژه‌های صحیح برای خطاهای غیرواژه و واژه حقیقی است. در مراحل شناسایی و پیشنهاد، معیار-F بهبود یافته است. ساختار معنایی مدل زبان پارسی اسپل و شبکه معنایی در شناسایی خطای واژه حقیقی استفاده شد. مقدار بازخوانی در شناسایی خطای واژه حقیقی به‌صورت معناداری بیشتر از نرم‌افزارهای رقیب آن است. نتایج ارزیابی نشان داده است که سامانه پارسی اسپل خطاهای واژه حقیقی بیشتری را شناسایی کرده و قادر است تا جایگزین‌های صحیح را برای واژه‌های نادرست پیشنهاد دهد. عملکرد کلی سامانه برای متون تخصصی رشته‌های مختلف بسیار مطلوب است. با این حال، از آنجا که مجموعه‌ای از واژگان متنی که در مدل زبانی چندوزنی استفاده شده، مجموعه بزرگی نیست، عملکرد کلی سامانه هنوز هم می‌تواند توسط مرمت مدل زبانی با استفاده از متون تک‌زبانه بیشتر در زمینه‌های مختلف موضوعی اصلاح شود.

۷- مراجع
7- References

[1] A. Sorokin, "Spelling Correction for Morphologically Rich language: a case study of Russian," in *Proceeding of the 6th on Balto-Slavic Natural Language Processing*, Valencia, Spain, pp. 45-53, 2017.

[2] K. Kukich, "Techniques for automatically correcting words in text", *ACM Computing Surveys (CSUR),* vol. 24, pp. 377–439, 1992.

[3] O. Kashefi, M Sharifi, and B. Minaie," A novel string distance metric for ranking Persian respelling suggestions", *Natural Language Engineering*, vol. 19, pp. 259–84, 2013.

[4] R. Mitton, "Ordering the suggestions of a spellchecker without using context", *Natural Language Engineering*, vol. 15, pp. 173–192, 2008.

[5] F. J. Damerau, "A technique for computer detection and correction of spelling errors", *Communications of the ACM*, vol.7, pp. 171–6, 1964.

[6] J. C. Wu, H. W ,Chiu, J. Chang, "Integrating dictionary and Web N-grams for chinese spell checking", *Computational Linguistics and Chinese Language Processing*, vol.18, pp.17–30, 2013.

[7] M. Janidarmian, A. Roshan Fekr, K. Radecka, Z. Zilic, "A comprehensive analysis on wearable acceleration sensors in human activity recognition", Sensors. vol 17, No. 3, 2017.

[8] N. Gupta and M. Pratistha, "Spell Checking Techniques in NLP: A Survey", *International Journal of Advanced Research in Computer Science and Software Engineering*, vol 2, Issue 12, December 2012.

[9] F. Ahmed and et al, "Revised N-Gram based Automatic Spelling Correction Tool to Improve Retrieval Effectiveness" [online].Available:http://-citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.186.3996

[10] D. Naber, "A Rule-Based Style and Grammar Checker", 2003, [online].Available: http://www-.danielnaber.de/languagetool/download/style_and_grammar_checker.pdf

[11] R. A. Wagner and M. J. Fischer, "The string-to-string correction problem," J. ACM, vol. 21, no. 1, pp. 168–173, 1974.

[12] E. Zamora, J. Pollock, "The use of trigram analysis for spelling error Detection", *Information Pro-cessing & Management*, vol 17, pp. 305-316, 1981.

[13] K. Toutanova and R. C. Moore, "Pronunciation modeling for improved spelling correction". In *Proceedings of the 40th Annual Meeting on. Association for Computational Linguistics*, pp. 144–151, 2002.

[14] J. Schaback and F. Li, "Multi-level feature extraction for spelling correction", in *IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data*, pp.79–86, 2007.

[DOI: 10.29252/jsdp.16.3.128]

[15] R. Mitton, "Spelling checkers, spelling correctors and the misspellings of poor spellers," Inf. Process. Manag., vol. 23, pp. 495–505, 1987.

[16] T. M. Miangah, "FarsiSpell: a spell-checking system for Persian using a large monolingual corpus". Literary and Linguistic Computing, vol 29, pp. 56–73, 2014.

[17] L. Barar, and B. QasemiZadeh,"CloniZER Spell Checker Adaptive Language Independent Spell Checker" In *AIML Conference CICC*, Cairo, Egypt, pp. 19–21, 2005.

[18] M. S. Rasooli, O.Kahefi, and B.Minaei-Bidgoli, "Effect of Adaptive Spell Checking in Persian" *in Natural Language Processing and Knowledge Engineering (NLP-KE), 7th International Conference on IEEE*, 2011. pp. 161–4.

[19] M. Shamsfard, H.S. Jafari, and M.Ilbeygi, "STeP-1: A Set of Fundamental Tools for Persian Text" in Processing. LREC, Malta, 2010.

[20] O, Kashefi, M. Nasri, and K. Kanani." Automatic Spell Checking in Persian Language". In *Supreme Council of Information and Communication Technology (SCICT)*, Tehran, Iran, 2010.

[21] H. Faili, N. Ehsan, M. Montazery and M. T. Pilehvar, "Vafa spell-checker for detecting spelling, grammatical,and real-word errors of Persian language," *Literary and Linguistic Computing*, vol. 31, pp. 95-117, 2016.

[22] M. Shamsfard, "Challenges and open problems in Persian text processing," Proc. LTC, vol. 11, 2011.

[23] P. Samanta and B. Chaudhuri, "A simple Readword Error Detection and Correction Using Local Word Bigram and Trigram," in Twenty-Fifth Conference on Computational Linguistics and Speech Processing (ROCLING 2013), Taiwan, R.O.C, 2013. pp. 211-220.

محمدباقر دستغیب، تحصیلات خود را در رشته مهندسی نرم‌افزار در مقطع کارشناسی در سال ۱۳۷۷در شیراز به پایان رساند و در مقطع کارشناسی ارشد هوش مصنوعی و رباتیک در دانشگاه شیراز تحصیلات خود را ادامه داد. وی در سال ۱۳۹۵ در گرایش هوش مصنوعی از دانشگاه شیراز موفق به اخذ درجهٔ دکترا شد. ایشان عضو هیأت علمی پژوهشگاه مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری است. زمینه‌های پژوهشی مورد علاقه ایشان، پردازش زبان طبیعی، بازیابی اطلاعات، پردازش تصویر است.

نشانی رایانامه ایشان عبارت است از:

Dastghaib@ricest.ac.ir

سارا کلینی کارشناسی ارشد هوش مصنوعی و رباتیک از دانشگاه شیراز، تألیف و ترجمه پنج کتاب در زمینه‌های بازیابی اطلاعات و فناوری اطلاعات. زمینهٔ تخصصی مورد علاقهٔ ایشان ذخیره و بازیابی اطلاعات است.

نشانی رایانامه ایشان عبارت است از:

koleini@ricest.ac.ir

سید مصطفی فخر احمد وی فارغ التحصیل دکترا از دانشگاه شیراز در گرایش نرم‌افزار است. ایشان هم اکنون عضو هیأت علمی دانشکده برق و کامپیوتر دانشگاه شیراز هستند. زمینه‌های پژوهشی مورد علاقه ایشان ماشین ترجمه‌ها، الگوریتم‌های پردازش متن، بازیابی اطلاعات و متن‌کاوی است.

نشانی رایانامه ایشان عبارت است از:

fakhrahmad@shirazu.ac.ir

[DOI: 10.29252/jsdp.16.3.128]