نخستین کنفرانس بین المللی پردازش خط و زبان فارسی
۱۵ و ۱۶ شهریور ۱۳۹۱
دانشگاه سمنان - دانشکده مهندسی برق و کامپیوتر

پیاده سازی یک غلط یاب املایی فارسی تحت وب

احمد یوسفان[۱] و بی بی صدیقه طباطبایی[۲]
[۱] کاشان، بلوار قطب راوندی، دانشگاه کاشان، دانشکده مهندسی، گروه مهندسی کامپیوتر، گروه مهندسی کامپیوتر، yoosofan@kashanu.ac.ir
[۲] دانش آموخته‌ی گروه مهندسی کامپیوتر دانشگاه کاشان، tabatabaeised@gmail.com

چکیده

غلطیاب املایی فارسی یکی از ابزارهای مهمی است که در راستای کمک به نویسنده‌ی یک متن فارسی می‌تواند به او کمک شایانی در یافتن و درست کردن غلط فارسی نوشته شده در یک متن نماید. تا کنون غلطیاب‌های گوناگونی آماده شده و در برخی از ابزارهای نگارش به کار گرفته شده است. با این همه در این زمینه پژوهش هنوز ادامه دارد زیرا کارآیی بسیاری از این ابزارها در حد بالایی نیست و باید الگوریتم‌های تازه‌ای برای بهبود آنها پیشنهاد شود. به کارگیری ایده‌های نو در کنار دیگر کارهای پیشین می‌تواند به بهبود کار غلطیاب‌های کنونی کمک شایانی نموده و برخی از مشکل‌های آنها را برطرف نماید. در این مقاله در آغاز برخی از روش‌های متداول غلط‌یابی بررسی شده است سپس روشی ترکیبی برای غلطیاب املایی فارسی پیشنهاد شده و این پیشنهاد به کمک زبان javascript و php و html پیاده سازی شده است. تحت وب بودن این ابزار پیاده سازی شده کمک می‌کند تا بتوان به سادگی آن را آزمایش نمود و اشکال‌های آن را برطرف نمود.

کلمات کلیدی

غلطیاب املایی، فارسی، پیاده سازی، وب، php، javascript

۱- مقدمه

زبان فارسی دربردارنده گنجینه‌ی بزرگی از زیباترین سروده‌ها و داستانها است. زبان فارسی یکی از پربارترین زبان‌های دنیا است. کتاب‌هایی چون مثنوی معنوی، دیوان حافظ، رباعیّات خیام و ... به زبان‌های گوناگون گیتی برگردانده شده و بارها چاپ شده‌اند. برترین ویژگی این نوشته‌ها، انسانی بودن آنها است بگونه‌ای که همه‌ی انسان‌ها گرایشی درونی به این نوشته‌ها دارند. متأسفانه این درخت تنومند امروزه نیاز به توجه بیشتری دارد زیرا برای دنیای نوین آماده نشده است[۱].

این زبان که از نیمه‌ی سده‌ی سوم، آثاری از آن در دست است و از آن تاریخ به بعد، روز به روز گسترش یافته و آثار بی‌شماری در آن آفریده شده، در حیطه‌ی فرهنگ آسیایی و بلکه جهانی، تبدیل به زبانی شده است که شاهکارهای جهانی آفریده است؛ چنان که کمتر زبانی از زبان‌های زنده‌ی موجود می‌توان سراغ داشت که در این پهنه با زبان فارسی کوس برابری زند. از جمله ویژگی‌های این زبان ، دیرندگی آن است در طول بیش از هزار سال؛ بدین معنی که دگرگونی‌ها در آن، در مقام سنجش با بسیاری از زبان‌های جهان، به نسبت کمتر بوده است؛ به طوری که فارسی زبانان امروز به آسانی می‌توانند شعر فردوسی را بخوانند و بفهمند و از آن بالاتر این که در آثار بسیاری از گویندگان و نویسندگان سده‌های اخیر و حتی معاصر می‌توان کاربردهای کهن هزار سال پیش را دید؛ حتی در کتاب‌های دوره‌ی دبستانی می‌توان شعر رودکی و فردوسی را گنجانید[۲].

رواج صنعت چاپ و ماشین‌های تحریر و رایانه‌ها و شتاب زدگی در نوشتن، بی‌دقتی را در درست و زیبا نوشتن دامن زده است. همچنین داده‌ها و صفحه‌های روی شبکه‌ی جهانی که برخی باید روزانه تغییر کنند، این دردسر را چندین برابر کرده‌اند[۱].

متأسفانه بی‌دقتی و رواج بسیاری از خطاهای نگارشی به ویژه در صفحه‌های شخصی فارسی روی شبکه‌ی جهانی به اندازه‌ای زیاد شده است که بیم آن می‌رود زبان دیرپای فارسی دچار مشکلات گوناگونی گردد و نتواند جایگاه خود را در میان زبان‌های زنده‌ی دنیا حفظ نماید.

امروزه رایانه‌ها برای آماده کردن نوشته‌های گوناگون بسیار به کار برده می‌شوند و بیشتر کسانی که نوشته‌ای را آماده می‌کنند از ابزاری در رایانه برای آماده کردن نوشته‌ی خود کمک می‌گیرند. همانند دیگر زبان‌های دنیا باید ابزارهای گوناگونی به کمک نویسنده‌ی متن فارسی در دنیای کنونی بیاید تا بتواند به او در بهتر نوشتن متن فارسی و همچنین پرهیز از خطا کمک نمایند. همچنین افزون بر گزارش نادرست بودن یک کلمه پیشنهاد یا پیشنهادهایی برای کلمه‌ی جایگزین آن به کاربر بدهد تا او را در درست کردن آن کلمه کمک نماید.

ناهماهنگی‌های گوناگونی در نگارش خط فارسی دیده می‌شود همچنین در نگارش رایانه‌ای متن فارسی قالب‌ها، ابزارها و سیستم عامل‌های گوناگون و روش‌های گوناگون کد کردن نوشته‌ی فارسی دیده می‌شود که در [۷-۳ ، ۱] به برخی از آنها پرداخته شده است. فرهنگستان زبان و ادب فارسی کوشیده است برخی ناهماهنگی‌ها را در خط فارسی برطرف نماید و استاندارد یکسانی برای نگارش خط فارسی پیشنهاد دهد در آدرس persianacademy.ir به روزترین نسخه از این استاندارد با نام دستور خط فارسی گذاشته شده است گرچه بسیاری از نویسندگان و حتی برخی کتاب‌های درسی به خوبی این استاندارد را رعایت نمی‌کنند ولی به هر حال می‌تواند به عنوان پایه‌ای برای درست نویسی متن فارسی به کار گرفته شود.

تا کنون غلطیاب‌های گوناگونی برای زبان فارسی آماده شده است و پروژه‌های گوناگونی در این زمینه انجام شده است که در این میان می‌توان به ابزار آماده شده برای خطایابی فارسی از Microsoft

office 2003 تا ۲۰۱۰، خطایاب فارسی شورای عالی اطلاع رسانی اشاره نمود. پروژه‌های گوناگون دیگری نیز در این زمینه انجام شده است که شرکت‌های گوناگونی در ایران آنها را آماده نموده‌اند. همچنین خطایاب‌های متن‌بازی چون ispell، aspell و hunspell تا کنون آماده شده و به روز شده‌اند که در آنها زبان فارسی نیز افزوده شده است. با این همه این ابزارها هنوز کاستی‌های گوناگونی دارند و باید این کاستی‌ها به کمک پیشنهادها و پژوهش‌های تازه‌ای برطرف شود یا از مشکلاتِ آن‌ها کاسته شود. بنابراین کار پژوهشی هر چه بیشتر در این زمینه، به بهبود خطایاب‌های کنونی کمک می‌کند و همچنین علاقه‌مندان به این زمینه را با الگوریتم‌های غلطیابی آشنا می‌نماید. در این مقاله کوشش شده است که یک روش ترکیبی برای بهبود غلطیابی فارسی پیشنهاد شود و سپس به صورت تحت وب آماده شود.

۲- غلط املایی و انواع آن

غلط‌های املایی به دو دلیل ممکن است به وجود آیند[۸]؛ دلیل اول این است که در هنگام تایپ متن ممکن است خطایی اتفاق بیفتد مانند کلمه taht که حاصل از خطای تایپی بر روی کلمه that است و این گونه خطا حاصل از خطا در نوشتن کلمه است. دلیل دوم این است که خطای رخ داده ناشی از ندانستن شکل صحیح املای کلمه است در این موارد می‌توان کلمه seprate را مثال زد که حاصل از ندانستن املای صحیح کلمه separate است.

دو نوع خطای املایی در متن‌ها ظاهر می‌شود[۹]. خطاهای non-word و خطاهای real-word است. خطاهای non-word خطاهایی هستند که شکل غلط آن‌ها در زبان وجود ندارد مانند غلط املایی teh که خطای املایی کلمه the است. خطاهای real-word خطاهایی هستند که تنها با توجه به جمله می‌توان آنها را شناسایی کرد و شکل غلط آنها نیز در زبان وجود دارد برای نمونه form که به صورت from نوشته شود به طوری که هر دو شکل کلمه در زبان وجود دارد هم from و هم form در این نمونه باید در جمله مورد نظر کلمه form نوشته می‌شد ولی from نوشته شده است. این گونه خطاها کاملاً وابسته به جمله است.

در ادامه منظور از هر گونه خطای املایی، خطای non-word است، زیرا در این مقاله فقط به این نوع از خطا پرداخته شده است. خطاهای real-word مستلزم پردازش جمله و متن به صورت کامل و در صورت نیاز درک کامل آن از سوی برنامه است. هنوز راه درازی تا انجام این کار در زبان فارسی وجود دارد و غلطیاب‌های کنونی هیچکدام این نوع خطاها را به خوبی شناسایی نمی‌کنند.

انواع اصلی خطاها را می‌توان به صورت زیر ذکر کرد.

- خطاهایی که حاصل از اضافه شدن یک حرف است مانند acress که می‌تواند خطایی در نوشتن کلمه cress باشد.
- خطاهایی که حاصل از حذف شدن یک حرف در کلمه است، مانند acress که می‌تواند خطایی در نوشتن کلمه actress باشد.
- خطاهایی که حاصل از عوض شدن یک حرف با حرف دیگر است، مانند acress که می‌تواند خطایی در نوشتن کلمه across باشد. که در این نمونه حرف o به حرف e تبدیل شده است.
- خطاهای جابه جایی، یک حرف کلمه با یک حرف دیگر از همان کلمه جابه جا می‌شود؛ مانند acress که می‌تواند خطایی در نوشتن کلمه caress باشد، که در این نمونه دو حرف c و a جابه جا شده اند.

نتایج آماری از این نوع خطاها که برای زبان انگلیسی انجام شده است در جدول ۱ یکی از این نتایج نشان داده شده است[۸].

جدول (1): نتیجه‌های آماری از چهار نوع خطای اصلی

نوع خطا	لغت‌نامه Webster	سندهای دولتی
جابجایی	۴۷ (۱۳/۱٪)	۴ (۲/۶٪)
اضافه	۷۳ (۲۰/۳٪)	۲۹ (۱۸/۷٪)
حذف	۱۲۴ (۳۴/۴٪)	۴۹ (۳۱/۶٪)
جانشینی	۹۷ (۲۹/۶٪)	۶۲ (۴۰.۰٪)
کل	۳۴۱ (۹۴/۷٪)	۱۴۴ (۹۲/۹٪)

عامل‌هایی که در نمونه خطا مؤثر هستند عبارتند از:

- همسایگی حروف در صفحه کلید که منجر به خطاهای اساسی می‌شود.
- کاراکترهایی که برای تایپ آنها نیاز به فشردن کلید شیفت است.
- هم صدایی مانند اینکه به جای کلمه receive، کلمه recieve تایپ شود که هر دو آهنگی شبیه به هم دارند.
- شکل‌های شبیه به هم مانند اینکه به جای کلمه انتخاب، کلمه انتحاب تایپ شود.

۳- شناسایی خطا

شناسایی خطا شامل مراحل زیر است.

۱. شناسایی محدوده کلمه در متن باید در ابتدا صورت گیرد. عموما کاراکتر فاصله و نقطه محدوده کلمه را در متن مشخص می‌کنند. البته شناسایی محدوده کلمه در زبان فارسی با دشواری‌های فراوانی روبرو است که در [۱، ۳-۴ و ۱۰] توضیح داده شده است. در این زمینه پژوهش‌ها و پیشنهادهای گوناگونی انجام شده است که یکی از آنها در [۱۱] آمده است و در [۱۰] همین روش با تغییرها و ساده‌سازی هایی برای ریشه‌یابی فارسی به کار گرفته شده است.

۲. خطاهای real-word به راحتی نظیر کردن کلمه با کلمات لغت‌نامه قابل شناسایی نیست بلکه نیاز به اطلاعاتی در رابطه با متن می‌باشد.

خطاهای non-word به وسیله چک کردن کلمه در لغت‌نامه قابل شناسایی است. این روش دارای مشکلاتی است که عبارتند از:

- نگهداری به هنگام یک چنین لغت‌نامه‌ای دشوار است.
- گسترش لغت‌نامه به گونه‌ای که همه کلمات موجود در متن را پوشش دهد.
- زمان پاسخ گویی سیستم پایین نگه داشته شود.

یک راه حل سریعتر ولی غیر قابل اعتماد برای بررسی کردن متن از خطاها این است که کنترل شود که همه n-gramهای موجود در یک کلمه به ترکیبات ماندنی n-gram تعلق داشته باشند. همچنین می‌توان به کمک یک ریشه‌یاب خوب لغت‌نامه کوچکتری را به کار برد تا کارآیی افزایش یابد و فضای کمتری نیز نیاز باشد.

۴- تصحیح خطا

تصحیح خطا می‌تواند به دو صورت خودکار یا فعل و انفعالی انجام شود. تصحیح خطای خودکار به این صورت است که به صورت

خودکار کلمه غلط با کلمه درست جایگزین می‌شود بدون آنکه کاربر مداخله‌ای داشته باشد. بنابراین غلط یاب باید این توانایی را داشته باشد که بهترین کلمه درست پیشنهادی را به دست آورد. تصحیح خطای فعل و انفعالی نیز بدین صورت است که کاربر به صورت فعل و انفعالی می‌تواند یکی از پیشنهادهایی که برای تصحیح کلمه غلط داده شده است را خود انتخاب کند. بنابر این غلط یاب می‌تواند چندین پیشنهاد را برای تصحیح کلمه غلط به کاربر بدهد.

تصحیح کلمات بدون در نظر گرفتن متنی که درونش هستند شامل سه مسئله می‌شود[۱۲]:

الف- شناسایی خطا

ب- ساخت پیشنهادهایی برای تصحیح

ج- رتبه‌بندی پیشنهادها.

شناسایی خطا پیش از این شرح داده شد. ماتریس‌های n-gram در یک متن می‌توانند پیشنهادهای تصحیح و رتبه‌بندی را انجام می‌دهند. در اینجا اشاره‌گری نیاز است که ارتباط میان یک عنصر ماتریس n-gram را با داده‌ای که آن غنصر n-gram درون آن است برقرار نماید. برای داده‌های زیاد (تعداد کلمه‌ها و اسناد زیاد) این راه حل عملی به نظر نمی‌رسد. اگر تعداد کلمه‌ها زیاد باشد یک روش برای کاهش تعدادِ کلمه‌های پیشنهاد شده برای تصحیح این است که زیر مجموعه‌ای از کلمه‌ها جستجو شود. با این کار احتمال از دست دادن پیشنهاد کلمه‌ی درست وجود دارد. بنابراین باید بر روی هر الگوریتم پیشنهادی دو عامل زمان پاسخ‌گویی و کیفیت پاسخ بررسی انجام شود.

معمولاً رتبه‌بندی میزان شباهت میان کلمه‌ی مورد بررسی و کلمه‌ی پیشنهادی را می‌سنجد و برای این کار از نتیجه‌ی الگوریتم شباهت کمک می‌گیرد. رتبه‌بندی می‌تواند ضعیف یا قوی باشد. در رتبه‌بندی ضعیف کلمه‌های پیشنهادی می‌توانند رتبه‌ی یکسانی داشته باشند و بنابراین شایستگی یکسانی را برای جانشینی دارند. افراطی‌ترین گونه‌ی رتبه‌بندی ضعیف آن است که همه‌ی کلمه‌های پیشنهادی رتبه‌ی یکسانی داشته باشند یا به عبارت بهتر اصلاً رتبه‌گذاری انجام نشود و فقط از میان کلمه‌های لغت‌نامه یا کلمه‌ای پیشنهاد شود یا خیر (تصمیم‌گیری دودویی). در رتبه‌بندی قوی به هر کلمه‌ی پیشنهادی یک رتبه‌ی یکتا نسبت داده می‌شود. در برنامه‌های تصحیحِ خودکار رتبه‌بندی قوی مورد نیاز است.

۵- روش‌های بررسی شباهت میان کلمات

در اینجا باید بتوان کلمه‌ها را با هم تطبیق داد؛ بنابراین به روش‌هایی برای بررسی و شباهت و نزدیکی کلمات و کلید شباهت پرداخته می‌شود. در این قسمت الگوریتم‌های متفاوتی تقسیم بندی شده‌اند. که همه یک حقیقت مشترک را دارند و آن این است که آنها سعی دارند که معیاری را برای فاصله میان کلمات بدهند. این معیارهای فاصله می‌تواند انواع زیادی داشته باشند.

۵-۱- معیار فاصله Damereau

چهار نوع از خطاهای املایی تک‌کاراکتری رایج‌تر در متون تایپ شده با ماشین را در [۹] شناسایی شده و یک الگوریتم برای تصحیح آنها ارائه شده است. این خطاها اضافه، حذف، جانشینی و جابه‌جایی هستند و تخمین زده شد که ۸۰ درصد از خطاهای موجود در متن از این چهار نوع می‌باشند. بنابراین همه اضافه‌ها، حذف‌ها، جانشینی‌ها و جابه جایی‌های ممکن روی کلمه با املای غلط آزمایش می‌شود و سپس کلمات تغیر یافته با لغتنامه نظیر می‌شوند. این روش یک مجموعه‌ای از $k*(2*n + 1) + n - 1$ گزاره صحیح را ارائه می‌دهد. به طوری که k تعداد حروف الفبای زبان و n طول کلمه با املای غلط است. برای یک کلمه ۷ کاراکتری در زبان سوئدی (۲۹ حرف) یک مجموعه ۳۴۱ کلمه‌ای باید با لغتنامه بررسی شوند. اگر ترکیباتی از حروف که نمی‌توانند وجود داشته باشند، حذف شوند؛ نتیجه می‌تواند به نصف کاهش یابد. به طور مثال در [۱۳] فیلتر اعتباریابی letter-N-Gram به صورت موثری در این مورد کار کرد. هزینه‌های زمانی می‌تواند با الگوریتم‌های تصفیه شده حذف شود به عنوان نمونه Domeij شخصی بود که این را به وسیله استفاده از یک روش بسیار سریع برای جستجوی لغتنامه، به صورت کارآمد اجرا کرد. همچنین این روش به وسیله Zamora & Pollock با لغتنامه کاهیده‌شده سازگار شده است. این لغتنامه کاهیده‌شده به وسیله روش‌های کلید شباهت خودشان بدست آمده است.

اصل علمی که این رویه دنبال می‌کند کار مستقیم با کلمه غلط ، تولید تصحیحات ممکن و سپس مقایسه آنها با لغتنامه است. به این روش نام دیگر فاصله ویرایشی معکوس نیز داده اند. بیشتر روش‌های دیگری که وجود دارند این گونه کار می‌کنند که در ابتدا دو رشته را داده و سپس شباهت آنها را بررسی می‌کنند.

۵-۲- Damerau–Levenshtein

روشن است که راه حل Damereau خطاهای املایی را در کلمات با حداکثر یک خطا حل می‌کند. اصل Damereau از چهار خطای املایی متفاوت اضافه کردن، حذف، جانشینی، جابه‌جایی بهره می‌گیرد که برای مقایسه فاصله ویرایشی بین دو کلمه طراحی شده است. بنابراین یک پایگاه داده از نام‌های درست و کامل می‌تواند بر طبق شباهت و نزدیکی به کلمه ورودی رتبه‌بندی شود. Damerau–Levenshtein با یک خطای تنها یا به نوعی دیگر فاصله ویرایشی ۱ متوقف نمی‌شود. در عوض مستقل از اینکه تغییرات چه چیز را می‌سازد، کوتاهترین فاصله بین یک کلمه(s1) و کلمه دیگر(s2) را محاسبه می‌شود. اندازه این فاصله با فرمول بازگشتی زیر محاسبه می‌شود که d یک ماتریس دوبعدی است.

$d(0,0)= 0$

$d(i,j)= \min\{$

$d(i-1,j)+1$, اضافه کردن

$d(i,j-1)+1$, حذف کردن

$d(i-1,j-1)+c(i,j)$, جانشینی

$d(i-2,j-2)+c(i,j-1)+c(i-1,j)+1$ جابه‌جایی

$\}$

اگر i مین حرف از کلمه اول برابر j مین حرف از کلمه دوم باشد آن‌گاه مقدار c(i,j) برابر صفر و گرنه برابر یک خواهد بود. این یک فاصله یا متریک استاندارد است که اغلب به وسیله برنامه‌نویسی پویا قابل اجرا است. رویه‌های دیگر برای حل مشکلات منظور شده‌اند که به طور مثال می‌توان از نمونه ای که توسط Brodda یا نمونه دیگری که توسط Du & chan معرفی شده نام برد. برای بهبود سرعت اجرا می‌توان از یک معیار برای برش استفاده کرد، منظور، برشِ قطعِ روندِ کار است؛ که قادر است هرگاه خطاهای زیادی شناسایی شد فرایند مقایسه را ناتمام رها کند.

۶- روش پیشنهادی

طی بررسی‌هایی که، نه به صورت آماری بلکه از روی نتایج به دست آمده از دیگر زبان‌هایی که رسم‌الخط آنها شبیه فارسی است نظیر زبان اردو، انجام شد، می‌توان انواع خطاهایی که بیشتر در زبان فارسی اتفاق می‌افتاد را به شکل زیر معرفی کرد.

- خطاهایی که حاصل از شباهت شکل دیداری کلمات است.
- خطاهایی که به دلیل تاثیر کلید شیفت اتفاق می‌افتند.
- خطاهایی که حاصل از حذف فاصله است.

۲۵ درصد از خطاها مربوط به خطاهای Real-word می‌باشند. مانند اسلامی که به صورت غلط سلامی تایپ شود. روش‌هایی که برای طراحی این غلط یاب استفاده شده است، تکنیک‌های Soundex و Edit Distance یا همان معیار Damereau است. معکوس روش Single Edit Distance در این پژوهش به کار گرفته شده است. و برای بررسی کردن اعتبار رشته‌های تصحیح تولید شده از لغت‌نامه استفاده شده است، به گونه‌ای که رشته تولیدی اگر جزء کلمات لغت‌نامه باشد معتبر شناخته می‌شود. تغییراتی که در روش Soundex داده شده است به قرار زیر بوده است:

- حرف اول نیز کد می‌شود.
- تعداد گروه‌های حرفی افزایش یافته است.
- از کدهای هگزا دسیمال به جای کدهای دسیمال استفاده شده است.

ابتدا تمام پیشنهادهای ممکن را با اجرای روش فاصله ویرایشی معکوس که در بالا شرح داده شد، به دست آورده و سپس برای رتبه بندی، ابتدا پیشنهادهایی که از روش Edit Distance برای حذف فاصله مجازی به دست آمده‌اند، دارای بالاترین رتبه قرار داده و سپس مابقی این پیشنهادها را برای رتبه‌بندی به الگوریتم Soundex داده و آن پیشنهادهایی که از نظر این الگوریتم پذیرفته شده‌اند در رتبه دوم قرار داده و مابقی پیشنهادهایی که مانده است را در رتبه چهارم. از بین این پیشنهادهای رتبه بندی شده، ۱۰ مورد که دارای بیشترین رتبه می‌باشند، انتخاب کرده و به کاربر نمایش داده شده است.

دلیل انتخاب این روش‌ها این است که می‌خواستیم هم تاثیر هم‌صدایی‌ها را در ایجاد خطاهای املایی لحاظ کنیم و هم معیاری را برای میزان تفاوت دو کلمه داشته باشیم. شاید در نظر شما الگوریتم‌های Metaphone یا Editex این کار را به خوبی انجام می‌دهند ولی باید توجه داشته باشید که این الگوریتم‌ها در ابتدا به قوانین موجود در زبان می‌پردازند در صورتی که در زبان فارسی قوانین زیادی برای صدای کلمات نداریم.

۷- پیاده سازی تحت وب

پرونده‌های نوشته شده برای این غلط یاب ۴ پرونده است:

- پرونده pellChecker.html: این پرونده، یک صفحه است که در مرحله اول به کاربر نمایش داده می‌شود تا متن خود را در آن وارد کند.
- پرونده spellChecker.php: این پرونده متنی را که کاربر وارد کرده است را از spellChecker.html دریافت کرده و عملیات غلط یابی و پیشنهاد تصحیح را انجام داده و نتیجه را به کاربر نمایش می‌دهد.
- پرونده lib.php: این پرونده تابع‌هایی را دربردارد که برای غلط یابی و تصحیح خطای پرونده spellChecker.php به آن نیاز دارد.
- پرونده dictionary.txt: این پرونده شامل کلمات صحیحی است که به عنوان لغتنامه از آن استفاده می‌شود. این لغتنامه در بردارنده‌ی همه‌ی کلمه‌هایی است که در زبان فارسی درست هستند و چون ریشه‌یابی در این غلط یاب به کار برده می‌شود همه‌ی کلمه‌ها و مشتقات آن در این پرونده گذاشته شده است. این پرونده بر پایه‌ی کارهای پیشین آماده شد و به صورت خودکار همه‌ی مشتقات کلمه‌ها ساخته شده بود و با کمک کلمه‌های گرفته شده از روزنامه‌ها با فراوانی بالاتر کلمه‌های درست برگزیده شده بودند و به این ترتیب می‌توان مشتقات کلمه‌ها ساخته شد. از سوی دیگر برخی از ترکیب‌ها که درستی آن‌ها روشن است بدون بررسی در کلمه‌های درون روزنامه‌ها به این مجموعه از کلمه‌ها افزوده شدند. برای کاهش سرعت جستجو در این کلمه‌ها می‌توان آن‌ها را به صورت مرتب درون پرونده گذاشت و سپس به کمک جستجوی دودویی روی آن‌ها جستجو کرد تا سرعت به دلیل اندازه‌ی بزرگ این پرونده کاهش نیابد.

غلط یاب به این ترتیب کار می‌کند که می‌توان کلمه یا کلمه‌هایی را وارد کرد و غلط یاب کلمه‌های نادرست را نشان می‌دهد و به ترتیب پیشنهادهای برای آن کلمه را نیز نشان می‌دهد سپس کاربر می تواند از میان پیشنهادها که به ترتیب به او نشان داده می‌شود یکی را برگزیند. در جدول زیر پیشنهادهای داده شده برای تعدادی از کلمه‌های داده شده نشان داده شده است.

جدول (۲): خروجی خطایاب

کلمه	پیشنهادهای داده شده
صلام	سلام، صلاحم، صلایم، غلام، کلام، آلام، صنام، صلحم، صلیم، صلاح، لام، صلا
هال	
سما	سنا، سمع، جسما، رسما، سیما، سرما، سمها، سماق، سماک، هما، شما، بما، اما، نما، دما، سقا، سزا، سرا، سوا، سمج، سمش، سمت، سمم، ما، سم، سام
حه	صحه، حقه، حصه، حبه، چه، سه، به، له، نه، مه، که، زه، ره، ده، آه، وه، حج، حق، حس، حی، حب، حل، حک، حر، حد
توره	توره، دوره، طفره، خوره، غوره، شوره، سوره، اوره، کوره، پوره، طوقه، طورش، طوری، طورت، طورم، طره، طور

این به این معنی است که تنها کلمات با همان شکلی که هستند بررسی می شوند و در لغتنامه نیز همه کلمات وجود دارد؛ به طور مثال هم باید کلمه "خوردند" و هم کلمه "می خورند"، "خوردم" و "خوردید" و تمام ترکیبات خوردن در این لغتنامه وجود داشته باشد؛ و هر چه حجم کلمات لغتنامه بیشتر، زمان اجرای عملیات غلطیابی نیز افزایش داده می شود. بنابراین یکی از مواردی که اگر در آینده به آن پرداخته شود در بهبود کارایی این غلط یاب بسیار موثر است این است که قبل و بعد از عملیات غلط یابی روش‌های دیگر نیز اجرا شوند که وظیفه این روش‌ها ریشه یابی کلمه ورودی و پس از غلط یابی، پیشنهادهایی که مطرح می شود، حالات مختلفی که می تواند با این ریشه پیشنهادی به کاربر داده شود، داده شود. یکی دیگر از بهبودهای ممکن برای این غلط یاب، بررسی روش های سریع جستجو در پرونده است یا بررسی بهترین حالتی که بتوان لغتنامه را نگهداری کرد.

۸- نتیجه

زبان فارسی به عنوان زبان رسمی ایران به عنوان یکی از نمادهای فرهنگ ایران به حساب می آید. کتاب‌های نوشته شده‌ی گذشتگانمان به این زبان به خوبی نشان دهنده‌ی تعلق این آثار گرانبها به ایران و فرهنگ ایران است. با وجود زبان‌ها و لهجه‌های گوناگونی که در ایران هستند که نباید این زبان‌ها و لهجه‌ها از میان برود و باید از آنها نگهداری کرد و به کار برد ولی نگهداری و به کارگیری زبان فارسی به عنوان یکی از نمادهای یکپارچگی و وحدت ایران ارزش بالایی دارد و کمرنگ شدن زبان فارسی به هر دلیلی، خیانت به یکی از فرهنگ‌های کهن گیتی است.

این مقاله به یکی از مشکلات رو در روی زبان فارسی، خطا در نوشتار، پرداخته است. برخی از پژوهش‌های انجام شده در این زمینه بررسی شده است و به بهره‌گیری از کارهای پیشین و ترکیب و تغییر برخی از آنها خطایاب ساده‌ای تحت وب طراحی و پیاده سازی شده است. گرچه این خطایاب توانایی بالای ندارد ولی ایده‌های به کار رفته در آن می‌تواند در کارهای آینده به کار گرفته شود و یا کارهای دیگر را بهبود بخشد.

مراجع

[۱] یوسفان، احمد. *يك سيستم بازيابي اطلاعات متني براي زبان فارسي بر پايه نمايه گذاري معاني پنهان*. پایان نامه کارشناسی ارشد، شیراز : دانشگاه شیراز ، شهریور ۱۳۸۲.

[۲] احمدی گیوی۱۳۸۰، حسن. *دستور تاریخی فعل*. تهران: نشر قطره، ۱۳۸۰، جلد اول، ص ۲۷.

[۳] یوسفان، احمد، سمیه صالحی، بهروز مینایی بیدگلی. *دشواری‌های ریشه‌یابی فارسی و روشی برای ریشه‌یابی فعل‌های ساده فارسی*، دومین کارگاه پژوهشی زبان فارسی و رایانه، تهران: دانشگاه تهران ، تیرماه ۱۳۸۵.

[۴] یوسفان۱۳۸۴، احمد، و منصور ذوالقدری جهرمی، مهدی احمدی. *روش خودکار شناسایی واژه‌های پربسامد در زبان فارسی*. کنفرانس بین المللی اطلاعات و دانش (IKT2005)، تهران:امیرکبیر، خرداد ۱۳۸۴.

[۵] احمدي بيرجندي ، احمد (۱۳۶۸). *شيوه‌ي آموزش املاي فارسي و نگارش*. تهران: انتشارات وزارت فرهنگ و ارشاد اسلامي . چاپ اول بهار ۱۳۶۸.

[6] Kokabi,Mortaza. *The Newer, theWorse: the Status of Farsi Word Processing Softwares in. Iran*, Proceedings ELPUB2005 Conference on Electronic Publishing – Kath. Univ. Leuven – June 2005

[۷] یوسفان، احمد. *ساخت یک نرم افزار خطایاب املایی و ریشه یاب خودکار فارسی*. طرح پژوهشی داخلی در دانشگاه کاشان ، ۱۳۸۸.

[8] Kukich, Karen. *A Comparison of Some Novel and Traditional Lexical Distance Metrics for Spelling Correction*. In: Proceedings of INNC-90-Paris, Paris, France, July 1990, pp 309-313.

[9] F. J. Damerau. 1964. *A technique for computer detection and correction of spelling errors*. In Communications of the ACM, volume 7(3), pages 171–176.

[۱۰] شیخ زادگان،جواد مرکز هوشمند علائم. *ساخت یک نرم افزار خطایاب املایی و ریشه یاب خودکار فارسی*. طرح پژوهشی داخلی در دانشگاه کاشان ، ۱۳۸۸.

[11]

[12] Sannella, M. J., *Constraint Satisfaction and Debugging for Interactive User Interfaces*, Ph.D. Thesis, University of Washington, Seattle, WA, 1994.

[13] Zachman, John A., *"A Framework for Information Systems Architecture"*, IBM Systems Journal, Vol. 26, No. 3, 1987.

[14] Plamondon, R., Lorette, G., *"Automatic Signature Verification and Writer Identification - The State of the Art"*, Pattern Recognition, Vol. 22, pp. 107-131, 1989.

[15] Object Management Group. *Unified Modeling Language: Superstructure*, Version 2.0, ptc/03-07-06, July 2003, http://www.omg.org/cgi-bin/doc?ptc/2003-08-02.