یک سامانه خودکار جهت خطایابی املایی متون فارسی موجود در وب

فاطمه سلطانزاده
دانشجوی کارشناسی ارشد
گروه زبان شناسی رایانشی
دانشگاه صنعتی شریف
fatemeh.slt@gmail.com

بهرام وزیرنژاد
استادیار
گروه زبان شناسی رایانشی
دانشگاه صنعتی شریف
bahram@sharif.edu

چکیده:

در این مقاله سامانه‌ای هوشمند جهت خطایابی املایی متون فارسی معرفی شده است. هدف از ارائه این سامانه، پردازش متون فارسی موجود در وب و خطایابی املایی خودکار آن‌هاست. این سامانه قادر است خطاهای املایی متون فارسی را تشخیص دهد و برای اصلاح کلمات نادرست، فهرستی از واژگان درست محتمل را پیشنهاد نماید. سپس کاربران وب می‌توانند از میان فهرست واژگان پیشنهادی واژه موردنظر را انتخاب نماید. پس از انتخاب واژه مطلوب کاربر، واژه مذکور به طور خودکار جایگزین واژه نادرست خواهدشد. نتایج ارزیابی حاکی از این است که این سامانه از دقتی معادل با 96.42I% در شناسایی و تشخیص خطا برخوردار است.

کلیدواژه ها: خطایابی املایی خودکار ، زبان فارسی، پردازش زبان طبیعی ، فاصله دمرا-لونشتاین، ستاکیابی واژگان.

1. مقدمه:

در جهان امروز دنیای وب حجم گسترده‌ای از اطلاعات را در اختیار کاربران خود قرار می‌دهد. حجم قابل توجهی از این اطلاعات به اطلاعات زبانی تعلق دارد و به صورت متون الکترونیکی در محیط وب موجود است. از آنجا که سخنوران بسیاری از دیرباز تا به امروز به زبان فارسی تکلم می‌کنند، بسیاری از متون موجود در وب به زبان فارسی نگارش شده‌اند. بدیهی است که این حجم چشمگیر اطلاعات نیازمند به اصلاح و ویرایش است تا از عاری هرگونه خطای املایی شود و کاربران بتوانند با سهولت به نیازهای خود دست یابند.

از سوی دیگر در بسیاری از کاربردهای پردازش زبان طبیعی هم‌چون بازشناسی خودکار نویسه‌های نوری، موتورهای جستجوگر، خلاصه‌سازی و استخراج واژگان کلیدی متون، ترجمه ماشینی، انواع داده‌کاوی و متن کاوی، سامانه‌های تبدیل متن به گفتار، دسته‌بندی و خوشه‌بندی متون، سامانه‌های پرسش و پاسخ و بسیاری امور دیگر، حجم عظیمی از اطلاعات توسط رایانه به طور خودکار مورد بازبینی و تحلیل قرار می‌گیرد. مسلم است که در صورت عدم یکپارچگی در ساختار متون و وجود خطاهای املایی و نگارشی فراوان، پردازش خودکار متون فارسی با چالش‌های بسیاری رو به رو خواهدبود. بنابراین وجود سامانه‌ای جهت خطایابی املایی خودکار زبان فارسی نه تنها خوانش متون فارسی را برای کاربران وب

تسهیل می‌بخشد، بلکه می‌تواند این متون را برای کاربردهای مختلف زبان طبیعی آماده نماید. اما مسائلی چند جهت پردازش خط فارسی وجود دارد که کار خطایابی املایی زبان فارسی را به خصوص در محیط مجازی با دشواری‌هایی رو به رو می‌سازد. در ادامه به شرح این مشکلات می‌پردازیم.

خط فارسی نیز مانند سایر خطوط دنیا، ویژگی‌هایی دارد که پردازش آن توسط رایانه را با مشکلاتی مواجه می‌سازد. این ویژگی ذاتی از آنجا ناشی می‌شود که در زبان‌های طبیعی بسیاری از واژگان در ترکیب با واژگان دیگر ساخته می‌شوند. اگر در نگارش واژگانی که از این ترکیب‌ها به‌وجودآمده‌اند قواعد مشخصی رعایت نشود، واژگان حاصل ممکن است معنایی متفاوت از آنچه مورد انتظار است را پیدا کنند. از دگر سو، تفاوت‌هایی نیز میان زبان فارسی و سایر زبان‌های طبیعی وجود دارد. در حقیقت، نحوه ساخت واژگان و اتصال آن‌ها در فارسی، دسته دیگری از مشکلات را در پردازش رایانشی زبان فارسی به‌وجودآورده است. ازمهم‌ترین چالش‌ها در حوزهٔ پردازش متون الکترونیک مسائلی مانند فاصله‌گذاری و نیم‌فاصله‌گذاری بین کلمات، اعراب در زبان فارسی، حضور حروف عربی و غیر استانداردی مانند «ی» و«ک» و کلمات عربی تنوین دار در متون فارسی است. تفاوت‌هایی نیز میان گفتار و نوشتار زبان فارسی وجود دارد. صورت گفتاری و نوشتاری در زبان فارسی به لحاظ سبک کلام، صرف واژگان و نحو جمله بسیار

متفاوت است. با توجه به اینکه حجم چشگیری از اطلاعات به زبان فارسی در محیط وب در بسیاری از وبلاگ‌ها و سایت‌های فارسی زبان به شکل محاوره‌ای نگارش شده‌است، پردازش رایانشی این متون از چالش‌های مهم در پردازش خط فارسی تلقی می‌شود (کاشفی و دیگران 1389: 10).

مشکل مهم دیگر در دستور خط فارسی این است برخی حروف در زبان فارسی دارای صورت آوایی مشابه ولی صورت نوشتاری متفاوت هستند (به مانند «ذ»، «ز»، «ظ» و «ض»). مورد عکس این مشکل نیز در زبان فارسی وجود دارد، برای مثال واژهٔ «ورود» را در نظر بگیرید. در این واژه حرف «و» بازنمایی یکسان از دو آوای متفاوت ارائه می‌دهد که اولی هم‌خوان «\v\» و دیگری واکهٔ «\u\» است.

تا آن زمان که این‌چنین مشکلات و ابهاماتی در نگارش متون فارسی وجود دارد، پردازش خودکار زبان فارسی با آسیب‌ها و دشواری‌های بسیار روبه‌رو خواهدبود؛ به همین دلیل خطایابی خودکار متون امری مهم و حیاتی تلقی می‌شود. در ادامه به معرفی پژوهش‌های انجام یافته در این راستا می‌پردازیم.

مطالعات در زمینه خطایاب املایی از اوایل دهه شصت میلادی آغاز شد. پس از آن پژوهش‌های دیگری در این راستا با استفاده از روش‌های یادگیری ماشینی، خطایابی با استفاده از چندوزنی‌ها، کلیدهای مشابهت، کلیدهای آوایی، روش‌های خطایابی بدون واژه

نامه انجام‌یافته است (کاشفی و دیگران 1389: 92). در سال 1990 مدل کانال نویزی در غلط‌یابی خودکار معرفی‌شد. در این مدل کلیه واژگان زبان از کانال نویزی عبور داده‌می شوند و خروجی کانال با کلمه ورودی مقایسه و محتمل‌ترین واژگان در قالب فهرستی به کاربر ارائه می‌شود (ژرافسکی و مارتین، 2007: 43). در پژوهشی (شیخ‌الاسلام و دیگران، 2012) ازمدل کانال نویزی در طراحی یک خطایاب املایی برای زبان فارسی استفاده شده‌است. در پژوهشی دیگر (بریل و مور، 2000) مدل پیچیده‌تری از کانل نویزی را معرفی شده‌است. در این روش به جای اصلاح حرف به حرف واژگان زبان، زیررشته‌هایی از لغات انتخاب شده و با زیررشته مناسب جایگزین می‌شود. این روش برای اصلاح خطاهای واحد تا چندگانه مفیدخواهد بود. در پژوهشی (فیلی و همکاران، 2010) یک خطایاب برای زبان فارسی ارائه‌شده‌است که توانایی تشخیص و تصحیح خطاهای املایی، نحوی و معنایی را داراست. خطایاب مذکور از یک روش مرتب سازی ترکیبی با استفاده از شباهت رشته‌ای و بسامد واژه استفاده می‌کند. در این روش جدولی شامل شباهت میان حروف زبان فارسی تولید شده‌است که با استفاده از آن شباهت میان رشته‌ها محاسبه‌می‌شود (کاشفی و دیگران 1389: 93). در طراحی خطایابی املایی خودکار دیگر برای ذخیره‌سازی واژگان فارسی، از الگوریتم سریع ساخت MADFA استفاده‌شده‌است. بدین‌ترتیب، علاوه بر افزایش سرعت، حجم

واژگان به طور متوسط به یک سوم کاهش‌یافته است. سپس از این مجموعه واژگان در خطایابی املایی استفاده‌شده است (دری نوگورانی و صبوریان، 1385).

دبیرخانه شورای عالی اطلاع رسانی (1390) محصولی را تحت عنوان «ویراستار» در جهت استانداردسازی متون و نویسه‌ها و خطایابی املایی واژگان فارسی تهیه‌کرده است. ویراستار به صورت متن‌باز در سرویس وب قابل دسترسی است. اما این محصول دارای کاستی‌هایی در پیش پردازش متون فارسی، همچون عدم درج نیم‌فاصله در جایگاه مناسب است.

در پژوهشی دیگر (شمس فرد و دیگران، 1388)، سامانه‌ای جهت پیش‌پردازش زبان فارسی شامل ویرایشگر، خطایاب املایی متون و تحلیل گر صرفی لغات زبان تهیه‌شده است. در خطایاب املایی مذکور، راهکاری برای تصحیح خطاهای ناشی از حذف فاصله بین کلمات ارائه شده است.

2. خطایاب خودکار املایی زبان فارسی:

خطایاب‌های املایی معمولا به تصحیح خطاهای حروف چینی، نگارشی و خطاهای ناشی از بازشناسی نویسه‌های نوری می‌پردازند. خطاهای حروفچینی به خطاهایی هم‌چون درج اشتباه نویسه‌های مجاور در صفحه کلید اشاره دارند. خطاهای نگارشی از عدم آگاهی نویسنده از قواعد و واژگان زبان ناشی می‌شود و خطاهای ناشی از بازشناسی

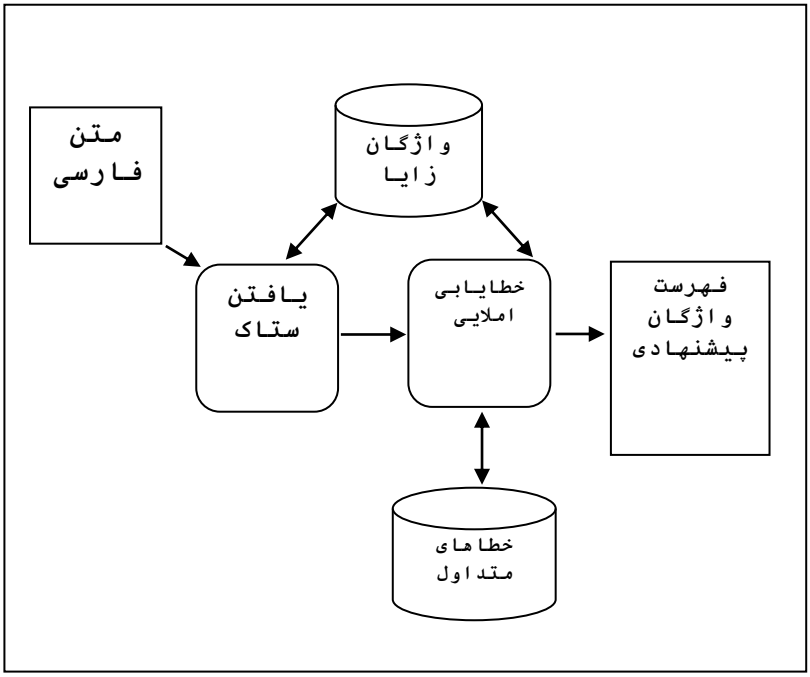
نویسه‌های نوری نیز به عدم موفقیت رایانه در بازشناسی یک نویسه خاص اشاره دارد. پژوهش‌های صورت گرفته بر روی پیکره‌های بزرگ نشان می‌دهد که 80% تا 90% از خطاهای املایی به دلیل چهار نوع خطای عمده روی می‌دهد. این چهار نوع خطا عبارتند از حذف یک حرف، درج یک حرف، جایگزینی یک حرف با حرف دیگر و جابه‌جایی دو حرف مجاور در واژهٔ صحیح. برای مثال واژهٔ «صلح» را در نظر بگیرید. این چهار نوع خطا برای واژهٔ مذکور در جدول زیر آمده است (کاشفی و دیگران، 1389 : 90)

جدول 1- انواع خطاهای املایی برای واژهٔ «صلح»

نوع خطا	مثال
درج	صلخح
حذف	صح
جایگزینی	صاح
جابه‌جایی	صحل

ما در این پژوهش این چهار نوع خطا را در خطایابی املایی تحت پوشش قرار داده‌ایم. فرآیند خطایابی در طی فرآیند زیر صورت می‌گیرد. ابتدا تمامی واژگان موجود در متن فارسی مورد بررسی قرار می‌گیرند و ستاک آنها استخراج می‌گردد. پس از آنکه ستاک واژه توسط ستاکیاب کشف و شناسایی شد، ستاک کلمه با تمامی واژگان موجود در

مجموعه واژگان زایا (اسلامی و دیگران، 1383) مقایسه می‌گردد. در صورت عدم وجود ستاک در واژگان زایا، برای واژهٔ مذکور مجموعه فهرستی از واژگان محتمل پیشنهاد می‌شود و برای برخی خطاهای متداول نیز به طور خودکار واژهٔ جدیدی جایگزین خواهدشد. معماری سیستم طراحی‌شده در این پژوهش را می‌توانید در شکل زیر مشاهده نمایید. در ادامه به شرح روش پیشنهادی می پردازیم.



شکل1 - معماری خطایاب املایی خودکار متون فارسی موجود در وب

این سامانه از یک روش ترکیبی در خطایابی املایی بهره می‌برد. در طی این پژوهش، فهرستی از واژگان نادرست متداول به همراه صورت صحیح آنها تدوین شده است. از آنجایی که این گونه از خطاها بسیار متداول است و همگان در صورت صحیح آن‌ها توافق نظر دارند، در صورت مواجهه با این نوع خطاها رایانه به طور خودکار واژه درست را جایگزین واژهٔ پیشین خواهد کرد.

روش دوم مورد استفاده در این خطایاب، روش فاصله دمرا-لونشتاین است.

DamerauLevenshtein(q, l)

{

$F_{dl}(0,0)=0$

If $(q_i==l_j)$*:*

$D(q_i,l_j)=0$

Else:

$D(q_i,l_j)=1$

If $((q_i==l_{j-1})$ *and* $(q_{i-1}==l_j))$*:*

$T(q_i,l_j)=0$

Else:

$T(q_i,l_j)=2$

$F_{dl}(i, j)= min((F_{dl}(i-1, j)+1),(F_{dl}(i, j-1)+1),(F_{dl}(i-1, j-1)+d(q_i,l_j)),(F_{dl}(i-2, j-2)+t(q_i,l_j)))$

Return $F_{dl}(|q|, |i|)$

شکل2- الگوریتم فاصله دمرا-لونشتاین

در صورتی که واژهٔ نادرست در این فهرست جایی نداشته باشد، با اجرای الگوریتم دمرا-لونشتاین نزدیکترین واژگان به واژهٔ نادرست یافته می‌شوند.طرح کلی الگوریتم در شکل 2 آمده است.

این الگوریتم دو واژه را به عنوان ورودی می‌گیرد و کوتاه‌ترین فاصله ممکن این دو واژه را محاسبه می‌نماید و به عنوان خروجی بازمی‌گرداند. در این روند زیررشته‌های متفاوت از دو واژه انتخاب‌شده و برای هر حالت انواع خطاهای احتمالی بررسی شده و حالتی با کوتاه‌ترین فاصله انتخاب می‌گردد. در این الگوریتم تشخیص چهار نوع خطای درج، حذف، جایگزینی و جابه‌جایی لحاظشده است. در اینجا واژگانی مانند «صح، صلخج، صاح» که دارای خطاهای حذف، درج و جایگزینی هستند دارای فاصله یک از واژهٔ «صلح» هستند و واژهٔ «صحل» که دارای غلط جابه‌جایی است، دارای فاصله دو از واژهٔ صحیح است. در روند خطایابی پژوهش حاضر، ابتدا فاصله واژهٔ نادرست از کلیه واژگان زایای زبان سنجیده‌شده‌است و سپس نتایج حاصل از آن ارزیابی شده‌است.

برای افزایش سرعت خطایابی املایی، محدودیت‌هایی در انتخاب واژگان برای راهیابی به فرآیند مقایسه کلمات لحاظ شده‌است که در ادامه به آنها اشاره خواهیم‌کرد.

یکی از محدودیت‌ها در انتخاب واژگان جهت استفاده در الگوریتم، مقایسه طول واژهٔ

محتمل با طول واژهٔ آغازین است. از آنجایی که در اکثر موارد اختلاف طول واژگان جایگزین کوچکتر از سه واحد است، در روند انتخاب واژگان تنها کلماتی بررسی می‌شوند که اختلاف طول آن‌ها در این بازه جای داشته‌باشد.

از دیگر محدودیت‌های لحاظشده می‌توان به برچسب اجزای کلام محتمل برای واژهٔ نادرست اشاره کرد. با توجه به ساختار تصریفی واژهٔ نادرست می‌توان نوع مقوله نحوی این واژه را حدس‌زد و تنها کلماتی را با آن مقایسه کرد که دارای مقوله نحوی یکسان با واژه آغازین باشند. پس از در نظرگرفتن محدودیت‌های انتخاب واژگان و محاسبه فاصله آن‌ها با کلمه آغازین، باید واژگان محتمل در قالب لیستی به ترتیب احتمال صحت واژه به کاربر پیشنهاد گردند. پس از آنکه کاربر یکی از واژگان فهرست را انتخاب کرد، آن واژه به طور خودکار به جای واژهٔ پیشین درج می‌شود. مرتب‌سازی این فهرست نیز بر اساس ترکیبی از فاصله واژه و مدل زبانی یونی‌گرم واژهٔ محتمل (که همان بسامد واژه در پیکرهٔ زبانی است) صورت می‌گیرد.

لازم به ذکر است که نسخه آزمایشی نرم افزار خطایاب املایی معرفی شده در این پژوهش در سرویس وب به طور رایگان و برخط قابل دسترس کلیه فارسی زبانان است[1].

[1] http://81.31.191.11/normalizer2/

در قسمت بعد به نحوهٔ ارزیابی این سامانه اشاره خواهیم‌کرد.

3. ارزیابی و گزارش نتایج:

در این قسمت به نحوهٔ ارزیابی سامانه و نتایج حاصل از آن می‌پردازیم.

برای ارزیابی سامانه خطایاب املایی خودکار از دو معیار دقت[1] و MRR[2] استفاده‌شده است. داده‌های آزمایشی دارای حجمی بالغ بر 500 واژهٔ زبان فارسی می‌باشند. نتایج ارزیابی حاکی از این است که این سامانه از دقتی معادل با 96.421% در شناسایی و تشخیص خطا برخوردار است. برای سنجش موفقیت سامانه در رتبه‌بندی فهرست واژگان جایگزین برای واژهٔ نادرست از معیار MRR استفاده‌کرده‌ایم. ارزیابی رتبه‌بندی این فهرست در دو حالت با لحاظ‌کردن محدودیت‌هایی مانند برچسب اجزای کلام محتمل برای واژهٔ نادرست و تفاوت طول دو واژهٔ درست ونادرست و بدون لحاظ‌کردن محدودیت‌های فوق انجام‌شده است و نتایج حاصل از این دو روش را با هم مقایسه‌شده است. در روش رتبه‌بندی بدون در نظر گرفتن محدودیت به MRR معادل با 0.4127 و در روش رتبه‌بندی با در نظر گرفتن محدودیت به MRR معادل با 0.6102 دست‌یافته‌ایم.

[1] Precision
[2] Mean Reciprocal Rank

4. نتیجه‌گیری:

نتایج ارزیابی این پژوهش بیانگر این است که سامانه معرفی شده به دقتی بسیار خوب در ویرایش و خطایابی املایی متون فارسی دستیافته است. همچنین درنظرگرفتن محدودیت‌های اختلاف طول و برچسب اجزای کلام واژگان علاوه بر افزایش سرعت منجر به رتبه‌بندی کاراتر شده است.

5. کارهای آتی:

به عنوان کارهای آتی می‌توان مدل زبانی مورد استفاده در سیستم را به مدل‌های زبانی مرتبه بالاتر بسط داد و از آن در جهت رفع انواع دیگری از خطایابی املایی بهره جست. بدین ترتیب خطایاب خواهد توانست خطاهایی که ناشی از استفاده از واژگان نابه‌جا در بافت متنی نامناسب هستند (برای مثال استفاده از واژه «حیات» به جای واژه «حیاط ») را نیز تشخیص داده و اصلاح نماید.

6. منابع:

کاشفی، امید، نصری، میترا، کنعانی، کامیار. (1389). خطایابی املایی خودکار در زبان فارسی؛ تهران، شورای عالی اطلاع‌رسانی، دبیرخانه.

صبوریان، محسن، دری نوگورانی، صادق. (1385). «طراحی و پیاده سازی یک خطایاب فارسی»؛ مجموعه مقالات دومین کارگاه

پژوهشی زبان فارسی و رایانه، دانشگاه تهران، تهران.

اسلامی، محرم، شریفی آتشگاه، مسعود، علیزاده لمجیری، صدیقه، زندی، طاهره. (1383). «واژگان زایای زبان فارسی»؛ مجموعه مقالات اولین کارگاه پژوهشی زبان فارسی و رایانه، دانشگاه تهران، تهران.

شمس‌فرد، مهرنوش، (1388). «Step1: تهیهٔ متن معیار برای زبان فارسی»؛ آزمایشگاه پردازش زبان طبیعی، دانشگاه شهید بهشتی، ایران.

H, Faili. (2010). “Detection and Correction of Real-Word Spelling Errors in Persian Language”. The 6th IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'10).

D, Jurafsky, J, Martin. (2007). Speech and language processing, An introduction to natural language processing, computational linguistics, and speech recognition, Prentice-Hall, Inc.

http://www.virastyar.ir/

M.H, Sheykholeslam, B, Minaei-Bidgoli and H, Juzi. (2012). “A Framework for Spelling Correction in Persian Language Using Noisy Channel Model”. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).

E, Brill , R. C, Moore. (2000). “An Improved Error Model for Noisy Channel Spelling Correction”. Proceedings of ACL.

جهت تماس با نویسندگان:

Tel: +98 21 66164833
Fax: +98 21 66029166