سامانه‌ای برای استانداردسازی و خطایابی متون علمی فارسی

هادی عبدی قویدل،[1] کارشناس ارشد زبان‌شناسی رایانشی، دانشگاه صنعتی شریف، تهران، ایران

ملوک‌السادات‌حسینی‌بهشتی،[2]* دکتری زبان‌شناسی همگانی، استادیار پژوهشگاه علوم و فناوری اطلاعات ایران، تهران، ایران

چکیده

روزانه هزاران مستند متنی متنوع در حوزه‌های مختلف علمی بر روی وب جهان‌گستر قرار می‌گیرد. این مستندات می‌تواند شامل پایان‌نامه‌ها، مقاله‌ها، گزارش‌های علمی و مواردی از این قبیل باشد. نگارش متن این مستندات علمی جهت حفظ یکنواختی باید بر اساس اصول ثابت انجام گیرد، اما همواره به طور غیر عمدی دست‌خوش سلیقه‌های مختلفی در طول تاریخ می‌شود. اگرچه این تغییرات ناشی از پویا بودن زبان و خلاقیت ذهن بشری است، اما این پویایی و خلاقیت پردازش ماشینی متن را با چالش‌های متعددی روبه‌رو می‌کند و دقت پردازش داده‌ها را به میزان چشمگیری پایین می‌آورد. علاوه بر تنوع نگارشی، غلط‌های سهوی املایی نیز وجود دارد که فحوای گفتمانی متن را منحرف کرده و درک آن را با مشکل مواجه می‌کند. بنابراین، کلیهٔ نویسه‌های متن باید به حالت استاندارد تبدیل شوند و عاری از هر گونه خطاهای املایی گردند. پژوهشگران مقالهٔ حاضر سامانه‌ای برای استانداردسازی و خطایابی متون علمی فارسی طراحی‌کرده‌اند که این سامانه متون نوشتاری علمی و تخصصی فارسی را به لحاظ صحت نگارشی و املایی بررسی می‌کند و متن را به شکل استاندارد در می‌آورد. در این مقاله، به معرفی کاربردهای سامانه می‌پردازیم.

کلیدواژه‌ها: مستندات علمی، پردازش ماشینی، صحت نگارشی، صحت املایی، شکل استاندارد

۱- مقدمه

خط فارسی بر اساس صادقی(۱۳۹۲) از خط عربی گرفته شده و خط عربی خود از خط فنیقی اقتباس شده و فنیقی مانند عربی متعلق به خانواده زبان‌های سامی است. وی معتقد است که در زبان‌های سامی، صامت‌ها اسکلت و پایهٔ معنایی کلمه را تشکیل می‌دهند و مصوت‌ها تنها برای گرفتن مشتقات مختلف از ریشه به کار می‌روند. چنانکه می‌دانیم فرهنگ‌های عربی

[1] **آدرس رایانامه:** habdi.cnlp@gmail.com

[2] **آدرس رایانامه:** beheshti@irandoc.ac.ir

نشانی: تهران ـ خیابان انقلاب ـ چهارراه فلسطین ـ ساختمان شماره ۱۰۹۰ ـ پژوهشگاه علوم و فنّاوری اطلاعات ایران ـ تهران ـ ایران

تقریبا همه بر اساس ریشهٔ کلمات تدوین شده‌اند. مثلا در عربی سه صامتِ «ف.ع.ل» به معنی «کردن» است. با افزودن دو فتحه یا دو مصوت **a** به این کلمه صورتِ «فَعَلْ» به وجود می‌آید که شکل ماضی این ریشه است. با افزودن یک فتحهٔ دیگر به پایان آن شکلِ «فَعَلَ» حاصل می‌شود که سوم شخص مفرد مذکر غایب این ماضی است.

صادقی معتقد است که زبان فارسی از دستهٔ زبان‌های خانوادهٔ هند و اروپایی محسوب می‌شود و ساختمان آن با ساخت عربی تفاوت‌هایی دارد. در زبان فارسی، صامت‌ها و مصوت‌ها پابه‌پای هم در ساختن صیغه‌های مختلف یک ریشهٔ مشارکت دارند. مثلا از ریشهٔ «دان» ما صیغه‌های «دانست، دانسته، داننده، دانا، نادان، می‌داند» و غیره را داریم که در ساختمان آن‌ها از مصوت‌ها و صامت‌های مختلف استفاده شده است. معنی کلمهٔ مرکب در زبان عربی و فارسی یکی نیست و شیوهٔ نگارش این نوع کلمات نیز در زبان عربی دارای مشکلات نگارشی زبان فارسی نیست. از طرف دیگر وجود دندانه و نقطه در خط عربی – فارسی خوانش کلمات را دچار مشکل می‌کند. مثلا خوانش نوشتن «زیست‌شناسی» به شکل «زیستشناسی» بسیار دشوار است.

به صورت کلی، مهمترین مشکلات فعلی خط فارسی از دیدگاه ذوالفقاری(۱۳۹۲) موارد زیر است:

– یکی نبودن موارد سرهم‌نویسی و جدانویسی و رواج شکل‌های گوناگون نوشتاری؛

– ابهام بسیار زیاد در خط فارسی به دلیل عدم تناظر میان حرف‌های الفبا و آواهای زبان؛

– آمیختن رسم‌الخط عربی با شیوهٔ خط فارسی و نگارش بی‌قاعدهٔ واژه‌ها و عبارت‌های برگرفته از عربی؛

ذوالفقاری همچنین بر این باور است که این مشکلات، منجر به پیدایش ناهماهنگی در نگارش موارد متعددی شده است که مهم‌ترین آن‌ها عبارت‌اند از:

– ترکیبات شامل «این» و «آن»؛
– پیشوند «ب» و حرف اضافه «به»؛
– پسوندهای «تر» و «ترین»؛
– پیشوندهای «بی»، «هم» و «هیچ»؛
– پیشوندهای فعلی «ب»، «ن»، «م»، «می» و «همی»؛
– علامت‌های جمع «ها» و «ان»؛
– فعل‌های ربطی یا استنادی: ام، ای، است، ایم، اید، اند؛
– ضمیرهای ملکی ومفعولی: م، ت، ش، مان، تان، شان؛
– کسره اضافه در حالت‌های مختلف؛
– برخی واژه‌های عربی‌تبار و عبارت‌های عربی رایج در فارسی؛

- اختلاف در نگارش همزه در حالت‌های مختلف...

چالش‌های پردازشی زبان فارسی را عبدی و بهشتی(۱۳۹۵) در ۱۰ دسته بر حسب هم‌آوایی، تأثیر حروف عربی بر متون زبان فارسی، ابهام یونی‌کد، چند املایی بودن، فاصله‌گذاری، شیوهٔ نویسه‌گردانی، دیدگاه صرفی و نحوی، پیوستگی حروفی و کاربرد اعراب طبقه‌بندی کرده‌اند. جدول ۱ مثال‌هایی را برای هرکدام ارائه می‌دهد.

جدول ۱ دسته‌بندی چالش‌های پردازشی زبان فارسی

دسته	مثال
هم‌آوایی	«ذ» و «ظ» و «ز» و «ض»
تأثیر حروف عربی بر متون زبان فارسی	«پاییز» و «پائیز» و یا مانند «حتما» و «حتماً»
ابهام یونی‌کد	«ی» و «ک»
چند املایی بودن	«بلیت» و «بلیط»
فاصله‌گذاری	«می‌خورد» گاهی به «می خورد» و «میخورد»
شیوهٔ نویسه‌گردانی	«abr» برای «اَبر»
دیدگاه صرفی	«آبشان»، «خوبان»
دیدگاه نحوی	«خانهٔ» و «خانه‌ی»
پیوستگی حروفی	«ما باهمدیگر با اتوبوس آمدیم»
کاربرد اعراب	«مَردَم»، «مَردُم» و «مُردَم»

سامانه‌های پیش‌پردازش مختلفی برای متون فارسی وجود دارد و تحقیقات بسیاری در قالب پژوهش آزمون و خطا انجام شده و یا در قالب محصول تجاری در بازار بین‌الملل عرضه شده است. این تحقیقات، هیچکدام به دقت ۱۰۰ درصد نرسیده‌اند و یا صرفا برای متون علمی فارسی طراحی نشده‌اند. بنابراین، در این پروژه سامانه‌ای طراحی شده است که بتواند:

- خطاهای نگارشی را در متن فارسی شناسایی و تصحیح کند.
- شیوهٔ نگارش را استاندارد کند.
- خطاهای املایی را در متن فارسی شناسایی و تصحیح کند.

۲- سامانهٔ استانداردساز و خطایاب متون علمی فارسی

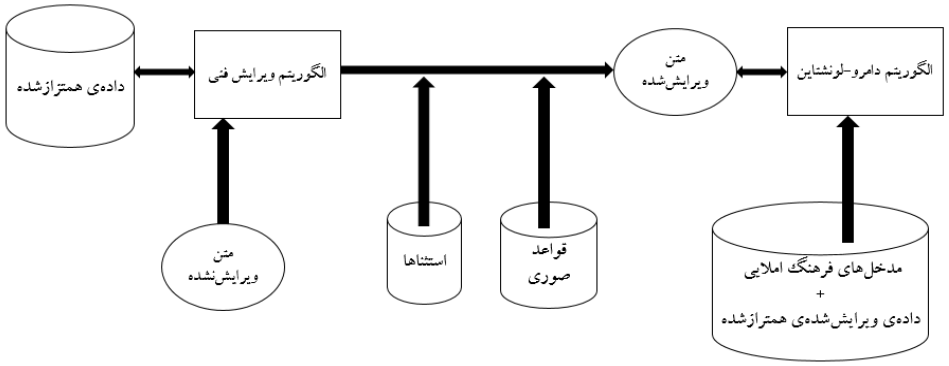
بی‌نظمی‌های موجود در خط فارسی و عدم رعایت یک‌دستی در نگارش، مشکلاتی را در ذخیره‌سازی استاندارد مدارک علمی و بازیابی ماشینی آنها ایجاد می‌کند. هدف از تولید سامانهٔ استانداردساز و خطایاب متون علمی فارسی، شناسایی، تصحیح و یکدست‌سازی خطاهای ویرایشی و نگارشی متون علمی فارسی است. برای طراحی چنین سامانه‌ای، پس از مرور پژوهش‌های نظری و پردازشی انجام شده و جمع‌بندی آنها، قواعد استانداردسازی متن که شامل نرمال‌سازی و خطایابی املایی می‌شود، استخراج شد. پس از آزمون اولیهٔ این قواعد، الگوریتم‌هایی برای آنها ساخته شده و پیاده‌سازی شد.

سامانهٔ استانداردساز و خطایاب متون علمی فارسی می‌تواند: خطاهای نگارشی را در متن فارسی شناسایی و تصحیح کند، شیوهٔ نگارش را استاندارد کند و خطاهای املایی را در متن فارسی شناسایی و تصحیح کند. عملکرد این سامانه به صورت مستقل در محیط وب، با قابلیت ارائهٔ سرویس به سامانه‌های دیگر، و جایگیری در نرم‌افزارهای دیگر می‌باشد. دادگان پژوهش حاضر بر دو نوع است: چکیده و مجموعهٔ کلمات. دادهٔ نوع اول حاوی ۹۰۰۰ چکیدهٔ مقاله و پایان‌نامه است. تعداد چکیدهٔ مختص هر حوزه به همراه میزان تنوع موضوعی آن در جدول ۲ ذکر شده‌است. دادهٔ نوع دوم حاوی ۲۹۱۰۵ کلمه از کتاب فرهنگ املایی(۱۳۹۱) استخراج شده است.

جدول ۲ حوزه‌ها و تعداد چکیده‌های مربوط به هر حوزه

حوزه	تعداد چکیده	میزان تنوع موضوعی
علوم انسانی	۴۰۰۰	۴۰۰
علوم پایه	۲۰۰۰	۲۲۰
فنی‌مهندسی	۳۰۰۰	۴۰۰
مجموع	۹۰۰۰	

فرایند کلی پردازش سامانه در شکل ۱ قابل ملاحظه است.



شکل ۱ فرایند پردازش متن توسط استانداردساز و خطایاب فارسی

۳- خروجی سامانهٔ استانداردساز و خطایاب متون علمی فارسی

سامانهٔ استانداردساز و خطایاب متون علمی فارسی عناصر متنی زیر را در دو مرحلهٔ استانداردسازی و خطایابی املایی به شکل متن استاندارد درمی‌آورد:

الف: استانداردسازی

- تاریخ: تاریخ‌ها در زبان فارسی به شکل‌های مختلفی نوشته می‌شود. مانند ۹۳٫۱٫۱ یا ۹۳/۱/۱ یا یکم فروردین ۱۳۹۳ و... . در فرایند استانداردسازی تمامی این فرمول‌ها تبدیل به یک فرمول واحد خواهند شد.

- حروف فارسی و عربی: تمامی حروف عربی طبق قوانین فرهنگستان زبان و ادب فارسی به حروف فارسی یکنواخت تغییر می‌یابند. مانند تبدیل «رئیس» به «رییس».

- فاصله و نیم‌فاصله: تمامی حالت‌های غلط فاصله و نیم‌فاصله تصحیح می‌شود. مانند تبدیل می شود و میشود به می‌شود. یا تبدیل نیم‌فاصله‌های غلط به نیم‌فاصله درست. مانند «می___شود» به «می‌شود»

- خط تیره: تبدیل خط تیره‌های انگلیسی به خط تیره‌های فارسی. «بند___یاب» به «بند-یاب»

- علایم نگارشی و کاربردهای متنوع هر نوع: تبدیل علایم نگارشی انگلیسی به فارسی(تبدیل «?» به «؟»)

- تبدیل واحدهای اندازه‌گیری مثل «کیلوگرم» و «kg» و یا «ک.گ.»

- اعداد(حروف و عددی): مانند تبدیل «یک» به «۱» و یا برعکس.

ب: تصحیح غلط‌های املایی:

واژه‌های پیشنهادی را برای کلمهٔ غلط ارائه می‌دهد. مانند: «دانش» برای «بانش»

شکل ۲ نمونه‌ای از جملات و ویرایش آنها را نشان‌می‌دهد.

شکل ۱ نمونه‌ای از ویرایش چکیده

٤- نتیجه‌گیری

مطالعهٔ حاضر نشان داد که بسیاری از چالش‌های موجود در زبان فارسی به طور کامل حل نشده و زمینهٔ تحقیق در مورد خطایابی و استانداردسازی متنی همچنان دارد. سامانهٔ حاضر نیز در میان سایر سامانه‌ها به استانداردسازی و خطایابی، این بار بر روی متون تخصصی و علمی نظیر مقاله‌ها و پایان‌نامه‌های فارسی، می‌پردازد. ارزیابی اولیهٔ $F\beta$ نشان داده است که این سامانه از دقت قابل توجهی برخوردار است.

٥- یادداشت

مقالهٔ حاضر برگرفته از طرح پژوهشی «طراحی و ساخت سامانهٔ استانداردساز و خطایاب متون فارسی» است که در پژوهشگاه علوم و فناوری اطلاعات ایران در حال انجام است.

فهرست منابع

ذوالفقاری، حسن،(۱۳۹۲). راهنمای ویراستاری و درست‌نویسی، چاپ سوم، تهران، نشر علم.

عبدی قویدل، هادی و حسینی‌بهشتی، ملوک‌السادات. ۱۳۹۵. پژوهشنامهٔ پردازش و مدیریت اطلاعات(زودآیند)

فرهنگستان زبان و ادب فارسی(۱۳۸۵). دستور خط فارسی. تهران، فرهنگستان زبان و ادب فارسی.

A System for Normalization and Spell Checking of Persian Scientific Texts

Hadi Abdi Ghavidel[۳]

Msc. Graduate in Computational Linguistics, Sharif University of Technology, Tehran, Iran

Molouk Sadat Hosseini Beheshti[٤]*****

PhD in General Linguistics; Assistant Professor; Iranian Research Institute for Information Science and Technology; Tehran, Iran

۳. habdi.cnlp@gmail.com.
٤. beheshti@irandoc.ac.ir.

Abstrcat

Every day, thousands of text documents in various scientific fields are placed on the World Wide Web. These documents may include theses, articles, scientific reports and etc. Writing the text of these scientific documents should be based on static principles in order to maintain uniformity, but the process is usually affected inadvertently by different tastes throughout the history. Although these changes result from dynamic nature of language and creativity of the human mind, this dynamism and creativity confront the machine processing of the texts with numerous challenges and considerably lower the accuracy of data processing. In addition to the writing variety, there are unintentional spelling errors which distort the discoursal content of the text and makes its difficult. Therefore, all the text characters should be converted into the normal form and become free from any spelling error. The researchers of the present paper have developed a system for normalization and spell checking of Persian scientific texts, which checks Persian written texts for the accuracy of its writing and spelling and converts the text into a normal form. In this paper, we introduce the applications of the system.

KeyWords :scientific documents, machine processing, accuracy of writing, accuracy of spelling, normal form