مجله پیاورد سلامت
دوره ۱۸، شماره ۱، فروردین ۱۴۰۳، ۳۱-۱۹
مقاله پژوهشی

تصحیح خودکار غلط‌های املایی در متون سونوگرافی فارسی با استفاده از شبکه‌های عصبی

سیدمحمدصادق دشتی[۱]، عمید خطیبی‌بردسیری[۲*]، مهدی جعفری‌شهباززاده[۳]

دریافت مقاله: ۱۴۰۲/۷/۱۷
پذیرش مقاله: ۱۴۰۲/۱۲/۲۵

* نویسنده مسئول:
عمید خطیبی بردسیری؛
دانشکده علوم پایه دانشگاه آزاد اسلامی واحد کرمان

Email:
a.khatibi@srbiau.ac.ir

چکیده

زمینه و هدف: گزارش‌های پزشکی و پرونده‌های الکترونیک سلامت برای تشخیص و درمان بیماران و تحقیقات پزشکی اهمیت فراوان دارند. تصحیح غلط‌های املایی موجود در متون پزشکی برای اطمینان از تفسیر صحیح اطلاعات امری ضروری است. این پژوهش برای تصحیح خودکار متون پزشکی زبان فارسی به کمک شبکه‌های عصبی انجام پذیرفته است.

روش بررسی: در این پژوهش که در سال ۱۴۰۲ انجام شد، مدل کامپیوتری جدیدی مبتنی بر شبکه‌های عصبی مصنوعی و تکنیک جای‌گذاری دوگانه با استفاده از زبان برنامه‌نویسی پایتون در محیط ویندوز توسعه یافت. مدل جای‌گذاری دوگانه کلمات به طور خاص برای تصحیح املا در حوزه متون سونوگرافی فارسی تنظیم شد. مدل پیشنهادی، از تکنیک‌های متنوعی برای تشخیص خودکار خطا، از جمله تطابق با فرهنگ واژگان ومحاسبه میزان مشابهت متنی استفاده می‌کند. همچنین برای انتخاب خودکار مناسب‌ترین کلمه جایگزین با غلط‌های املایی، از ویژگی‌های خاصی همچون فاصله ویرایش(Edit-Distance)، همراه با امتیاز مشابهت استفاده شده است. داده‌های آموزش و آزمایش مدل جاری، بخشی از مجموعه متون کلینیک سونوگرافی بیمارستان امام خمینی تهران است.

یافته‌ها: مدل پیشنهادی بر اساس شبکه‌های عصبی مصنوعی توسعه یافته و از یک معماری جدید جای‌گذاری دوگانه کلمات جهت انتخاب بهترین کلمات کاندید، به منظور جایگزینی با غلط‌های املایی و معنایی بهره می‌برد. مطابق بررسی انجام شده بر روی متون سونوگرافی فارسی، دقت مدل پیشنهادی بر حسب معیار F-Measure) F) در تشخیص و تصحیح خودکار خطاهای معنایی به ترتیب برابر با ۹۰/۵٪ و ۹۰٪ می‌باشد. به علاوه، دقت ۹۰/۸٪ در زمینه تصحیح خطاهای شکلی کسب گردید.

نتیجه‌گیری: مطابق نتایج ارزیابی، روش پیشنهادی می‌تواند به طور مؤثر طیف گسترده‌ای از خطاهای شکلی و معنایی، از جمله جایگزینی، جابه‌جایی، درج و حذف را در متون پزشکی مدیریت کند. استفاده و ادغام معیار فاصله ویرایش با امتیاز مشابهت متنی مستخرج از مدل جای‌گذاری دوگانه به‌طور قابل‌توجهی دقت تصحیح غلط‌های املایی را در متون سونوگرافی فارسی افزایش داده که این امر متضمن صحت بیش‌تر محتوای این گونه اسناد خواهد بود. به باور نویسندگان، مدل پیشنهادی، پیشرفت قابل‌توجهی در زمینه‌ی تشخیص و تصحیح غلط‌های املایی برای متون سونوگرافی زبان فارسی است.

واژه‌های کلیدی: تصحیح خطا، جای‌گذاری عصبی، شبکه‌های عصبی، متون سونوگرافی، پردازش زبان فارسی

۱ دکتری مهندسی کامپیوتر، دانشکده علوم پایه، واحد کرمان، دانشگاه آزاد اسلامی، کرمان، ایران

۲ استادیار گروه مهندسی کامپیوتر، دانشکده علوم پایه، واحد کرمان، دانشگاه آزاد اسلامی، کرمان، ایران

۳ استادیار گروه مهندسی برق، دانشکده فنی و مهندسی، واحد کرمان، دانشگاه آزاد اسلامی، کرمان، ایران

مقدمه

اصلاح املای کلمات یکی از فرایندهای مهم در تمام محیط‌های پردازش متن است؛ به خصوص برای زبان‌هایی با ساختار مورفولوژیک و دستور زبان پیچیده مانند زبان فارسی. این کار در حوزه‌ی متون بالینی حایز اهمیت بیش‌تری می‌باشد؛ زیرا که ثبت دقیق اسناد برای مراقبت از بیمار، تحقیق و تضمین ایمنی بیمار حایز اهمیت است(۱). کیفیت و ایمنی مراقبت‌های بهداشتی و درمانی، به دقت در ثبت اسناد بالینی بستگی دارد(۲). با این حال، اغلب در آماده‌سازی متونی از این دست، به دلیل فشار زمانی اشتباهات املایی رخ می‌دهد(۳).

فرایند اصلاح املا، عمدتاً دو نوع خطا را رفع می‌کند: خطاهای شکلی (Non-word error)، که کلمات بدون معنا و غیرموجود در فرهنگ لغت هستند، و خطاهای معنایی کلمات (Real-word error)، که کلمات با املای صحیح هستند اما خارج از زمینه بحث استفاده شده‌اند(٤). این خطاها می‌توانند از منابع مختلف نشأت گیرند، از جمله دلایل اشتباهات تایپی، می‌توان سردرگمی بین کلمات با صدا یا معنای مشابه، جایگزینی‌های نادرست توسط سامانه‌های خودکار مانند فعال بودن قابلیت تصحیح خودکار، و سوء تفسیر ورودی توسط سامانه‌های خودکار تشخیص صوت (Automatic Speech Recognition) و تشخیص کاراکترها (Optical Character Recognition) را برشمرد(٤).

طی سالیان اخیر مشخص گردیده است که به کارگیری اطلاعات معنایی (Semantic Information) کلمات در بهبود کارآمدی سامانه‌های خودکار اصلاح کننده‌ی متون مؤثر است(۵). در همین خصوص روش نوآورانه‌ای نیز برای اصلاح هم‌زمان چندین خطای شکلی در شرایط محیط پرخطا توسعه یافت(٦). همچنین دشتی مدلی را توسعه داد که در آن با تشخیص و خودکارسازی تصحیح خطای معنایی در مواردی که بیش از یک خطا در یک دنباله کلمات وجود دارد، پرداخته شده است(۷). روش‌های جدیدتر از اطلاعات زمینه‌ای (Context Information) با استفاده از شبکه‌های عصبی (Neural Networks) و به طور خاص جای‌گذاری کلمات (Word Embeddings) برای اصلاح املا استفاده کرده‌اند(۸). همچنین روش‌های یادگیری عمیق برای حل مشکل اصلاح خطای املایی در زبان‌های مختلف اعمال شده‌اند(۱۵-۹). در حوزه‌ی بهداشت و درمان، تکنیک‌های اصلاح املا در بسط مخفف‌ها و خلاصه‌سازی، و اصلاح اشتباهات املایی مؤثر بوده‌اند(۱۷و۱٦). حسب مشاهدات چنین خطاهایی تا ۳۰٪ محتوای بالینی را تشکیل می‌دهند(۱۸). همچنین چندین مدل قابل توجه در زمینه‌ی پزشکی برای اصلاح املای کلمات در متون فرانسوی(۱۹)، مجارستانی(۲۰)، متون بالینی سوئدی(۲۱) توسعه یافته است. از طرف دیگر، سیستم‌هایی نیز برای اصلاح خطاهای تایپی و زبانی در واژگان پزشکی (۲۲)، گزارش‌های ایمنی واکسن(۲۳)، گزارش‌های بالینی(٢٤)، اشتباهات نام دارو(۲۵) ارایه گردیده‌اند. دیگر مدل‌ها نیز شامل اصلاح خودکار املا برای متون پزشکی براساس مدل کانال نویزی (Noisy Channel Model)(۲٦)، روش بدون نظارت (Unsupervised Approach) برای اصلاح خطاهای املای معنایی در متون بالینی آزاد به زبان هلندی و انگلیسی(۲۷)، سامانه شناسایی و اصلاح خطاهای مربوط به سیستم تشخیص کاراکتر و مدل اصلاح خطا معنایی در متن بالینی می‌باشند(۲۹و۲۸).

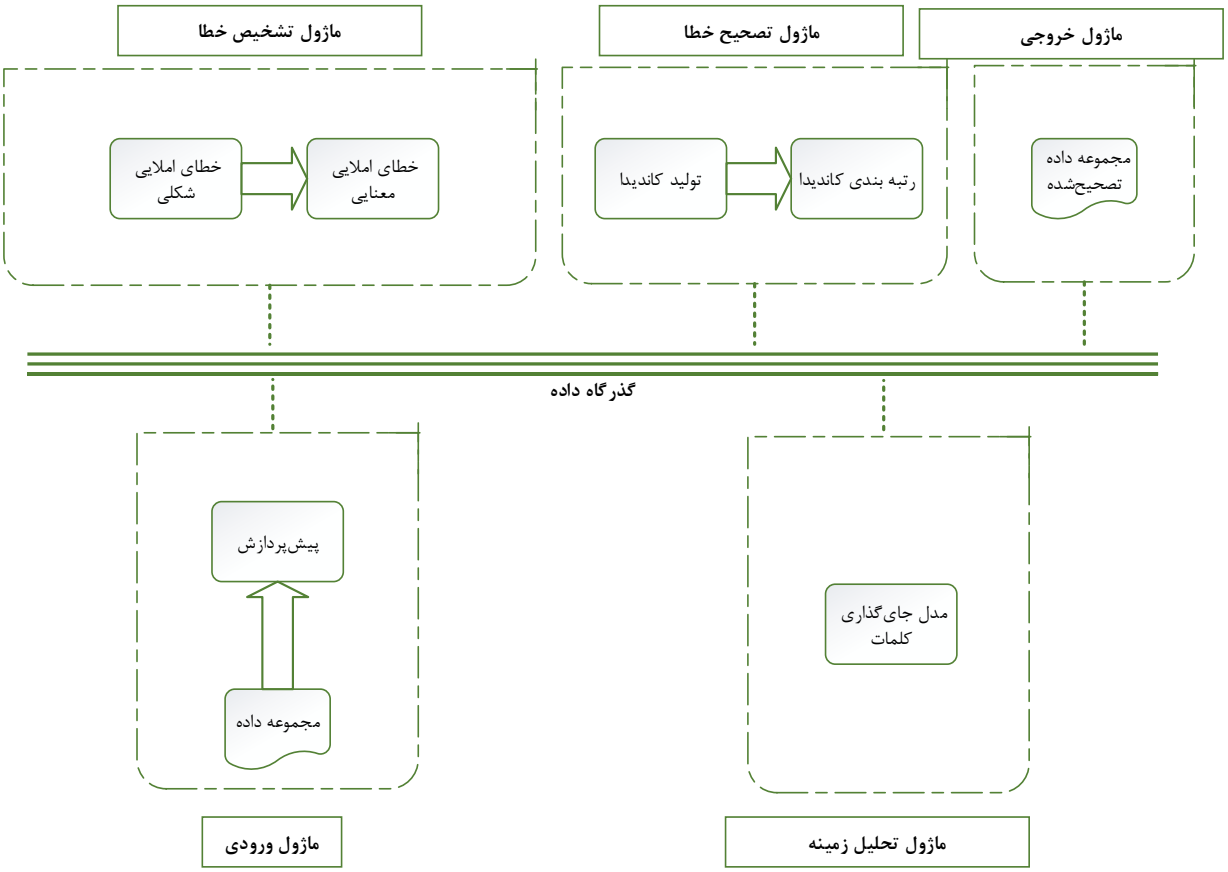
زبان فارسی با واژگان غنی و خصوصیات پیچیده‌ی خود، چالش‌هایی منحصر به فرد برای اصلاح خطای کلمات واقعی فرا هم می‌آورد. خصوصیات منحصر به فرد فارسی مانند همخوانی(کلماتی که صدایی یکسان دارند اما معانی مختلف)، چندینی(کلمات با چندین معنی)، هتروگرافی(کلمات با املای یکسان اما معانی متفاوت بسته به تلفظ)، و مسائل مربوط به حد فاصل کلمات به این پیچیدگی کمک می‌کنند(۳۰) .با وجود این چالش‌ها، پیشرفت قابل توجهی در حوزه‌ی اصلاح املای فارسی صورت گرفته است. روش‌های استفاده شده از روش‌های آماری یا قاعده‌محور تا سیستم‌های مدرن‌تر، مانند اصلاح کننده‌ی املای وفا، که قادر به شناسایی انواع گسترده‌ای از خطاهاست(۳۱). همچنین مدل ارایه‌شده توسط موسوی و میانگاه با استفاده ازمدل زبانی چندگرمی (N-grams)، یک مجموعه متن (Corpus) یک زبانه، و یک معیار فاصله رشته (String Distance)، مسایل املایی در زبان فارسی را برطرف کرده‌اند(٤۰-۳۲و۳۰). در بین این روش‌ها، تنها یکی به اصلاح اشتباهات املایی در متون سونوگرافی فارسی براساس مدل زبانی چهارگانه اختصاص یافته است(۳۹). با توجه به کمبود ابزارها و منابع پردازش زبان فارسی و با علم به این‌که این زبان نادر، مانند هر زبان دیگری ویژگی‌های خاص خود را دارد از جمله: حروف، معانی و نیز مرزبندی کلمات را دارا می‌باشد، نیاز به یک اصلاح‌کننده املای فارسی بالاخص در زمینه‌های تخصصی، از جمله بهداشت و درمان، آشکارا بیش از پیش برجسته می‌گردد(٤).

در این پژوهش، یک روش نوآورانه برای شناسایی و اصلاح خطاهای املایی در متون سونوگرافی فارسی معرفی می‌گردد. دستاورد اصلی این پژوهش یک مدل جای‌گذاری دوگانه‌ی کلمات است که به طور خاص برای اصلاح املا در حوزه‌ی سونوگرافی فارسی تنظیم شده و آموزش دیده است. همچنین روش

پیشنهادی با استفاده از معیارهای سنجش همچون دقت (Precision)، بازخوانی (Recall)، و معیار F- با سایر مدل‌های موجود در حوزه‌ی شناسایی خطاء املایی در متون رادیوگرافی فارسی مقایسه می‌گردد.

روش بررسی

پژوهش جاری از نوع کاربردی است که در سال ۱۴۰۲ بر روی مجموعه متون سونوگرافی فارسی سیستم اطلاعات سلامت بیمارستان امام خمینی تهران انجام شده است. مدل پیشنهادی دو نوع خطا را در متون سونوگرافی فارسی شناسایی و اصلاح می‌کند: خطاهای شکلی و خطاهای معنایی. معماری مدل پیشنهادی در شکل ۱ نشان داده شده است.



شکل ۱: معماری مدل پیشنهادی جهت تشخیص و تصحیح انواع خطا

مطابق شکل ۱، این مدل از پنج ماژول متمایز تشکیل شده است که از طریق یک گذرگاه داده (databus) ارتباط برقرار می‌کنند. ماژول ورودی، متن‌های خام را دریافت می‌کند. ماژول پیش‌پردازش (Preprocess) متن را نرمال می‌کند و مشکلات مربوط به فاصله درون کلمات را برطرف می‌کند. ماژول تحلیل‌زمینه (Context Analyzer)، سطح شباهت را درون توالی‌های کلمات مورد نظر ارزیابی می‌کند. برای تشخیص خطاهای شکلی، از روش جستجو در لیست واژگان فرهنگ لغت استفاده می‌گردد؛ اما برای تشخیص خطای معنایی، از آنالیز شباهت کلمات با زمینه بهره برده شده است. ماژول اصلاح خطا، هر دو دسته خطا را با استفاده از شباهت زمینه‌ای دریافتی از مدل جای‌گذاری کلمات اصلاح می‌کند. سپس، مجموعه های اصلاح شده یا توالی کلمات از طریق ماژول خروجی به کاربر نمایش داده می‌شود.

در این پژوهش از مجموعه متونی که توسط سیستم اطلاعات سلامت (Health Information System) بخش تصویربرداری بیمارستان امام خمینی در تهران فراهم شده است، استفاده گردید. این منابع داده، شامل سه نوع گزارش پزشکی مختلف بود: سونوگرافی پستان، سونوگرافی سر و گردن، و گزارش‌های سونوگرافی شکم و لگن. این گزارش‌ها توسط تایپیست‌های پزشکی بین دی‌ماه ۱۳۹۳ و خردادماه ۱۳۹۷ تولید شده‌اند؛ که در مجموع شامل ۳۷,۸۹۹ گزارش و ۱,۸۹۲,۷۳۴ کلمه می شود. از این مجموعه، به صورت تصادفی ۳۴,۶۴۲ گزارش و ۱,۷۴۵,۴۱۴ کلمه برای آموزش مدل دوگانه جای‌گذاری کلمات انتخاب گردید. باقی‌مانده ۳,۲۵۷ گزارش و ۱۴۷,۳۲۰ کلمه برای آزمایش و ارزیابی استفاده

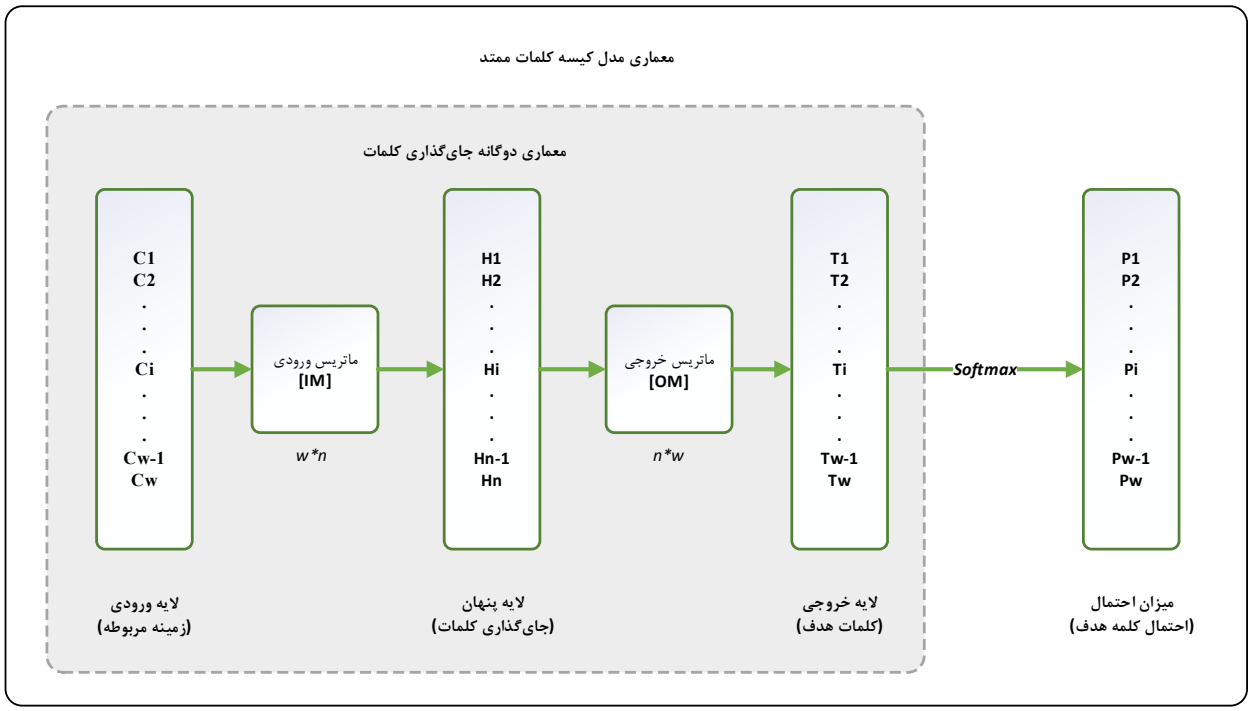
شدند. علاوه بر این، مجموعه‌های آموزش و آزمون مورد بررسی قرار گرفت و خطاهای شکلی و خطاهای معنایی را در هر دو مجموعه شناسایی شد. همچنین مشاهده شد که ٦,٥٨١ مقاله در مجموعه آموزش حداقل دارای یک یا بیش‌تر خطای شکلی هستند؛ که جمع آن‌ها به ١٥,٤٨٧ اشتباه شکلی بالغ می‌گردد. در همین حال، ١,٦٢٢ گزارش سونوگرافی حداقل دارای یک نمونه خطای معنایی بودند؛ که مجموعا برابر با ٣,٧٤٠ خطای معنایی است. تمام این خطاها قبل از آموزش مدل دوگانه جای‌گذاری کلمات، اصلاح شدند. مجموعه آزمون و تست شامل ٧٤٨ مقاله با حداقل یک خطای املایی در آن بود، که مجموعاً شامل ١,٧٢١ خطای شکلی می‌شد. علاوه بر این، ١٤٩ گزارش سونوگرافی حداقل دارای یک نمونه خطای معنایی بودند که برابر با ٣٥٤ خطای معنایی می‌باشد.

پیش‌پردازش متن، مرحله‌ی حیاتی در بیش‌تر برنامه‌های پردازش زبان طبیعی است؛ که شامل تقسیم جملات (Sentence Segmentation)، توکن‌سازی (Tokenization) و نرمال‌سازی است(٤). تقسیم جملات، فرایند شناسایی مرزهای جمله است که معمولاً با علایم نگارشی مانند نقطه، علامت تعجب یا علامت سوال مشخص می‌شوند. توکن‌سازی شامل تجزیه یک جمله به سری اصطلاحات است که معنای جمله را دربر می‌گیرد و برای استخراج ویژگی استفاده می‌شود. نرمال‌سازی، فرایند تبدیل متن به اشکال اصلی آن است و نقش قابل‌توجهی در برنامه‌های پردازش زبان طبیعی فارسی دارد، همان‌طور که در بسیاری از زبان های دیگر دارد. یک وظیفه‌ی کلیدی در نرمال‌سازی متن فارسی، تبدیل فرم پیوسته و فاصله به اشکال منظم و همچنین جایگزین کردن فاصله با نیم‌فاصله در صورت لزوم است. برای مثال، «می‌شود» با «می‌شود» جایگزین می‌شود. زبان‌های فارسی و عربی دارای بسیاری از خصوصیات مشترک هستند و بعضی از حروف فارسی اغلب با اشکال عربی اشتباه نوشته می‌شوند. پژوهشگران، اغلب موارد، اشتباهاتی ازین دست را با استاندارد سازی و جایگزین کردن حروف با فرم نوشتار عربی مانند: «ی، ک، ه » با معادل‌های فارسی آن‌ها برطرف می‌کنند. به عنوان مثال، «برائ» به «برای» تغییر می‌یابد. نرمال‌سازی شامل حذف نشانگان از کلمات فارسی نیز می‌شود؛ به عنوان نمونه، «ذرّه» به «ذره» تغییر می‌یابد. علاوه بر این، کشیده نیز از کلمات حذف می‌شود؛ به عنوان مثال، «بـــــــاند» به «باند» تغییر پیدا می‌کند. برای دستیابی به هدف نرمال‌سازی، از لغت نامه دهخدا به عنوان مرجع استفاده می‌شود که شکل املای صحیح تمام کلمات فارسی را دربردارد(٤١).

مدل پیشنهادی از معیار فاصله دامارو–لونشتین (Damerau-Levenshtein) برای تولید کاندید برای خطاهای شکلی و معنایی استفاده می‌کند(٤٢). معیار فاصله فوق چهار عملیات درج (Insertion)، حذف (Deletion)، جایگزینی (Substitution) و جابجایی (Transposition) حروف را در نظر می‌گیرد . به عنوان مثال، فاصله دامارو–لونشتین بین «نی» و «بین» ٢ واحد است. مشخص شده است که حدود ٨٠٪ از خطاهای املایی تولید شده توسط انسان این چهار نوع خطا را در بر می‌گیرد(٤٣). مطالعات نشان می‌دهد که خطاهای واقعی حسابی حدود ٢٥٪ تا ٤٠٪ از تمام خطاهای املایی در متن انگلیسی را تشکیل می‌دهند (٤٥و٤٤). مدل پیشنهادی از یک فرهنگ لغت جامع برای شناسایی اشتباهات استفاده می‌کند. این فرهنگ لغت به دو بخش عمومی و تخصصی تقسیم شده است. برای بخش عمومی، فرهنگ لغت استفاده شده در نرم‌افزار تصحیح خطای وفا استفاده گردید(٣١). این فرهنگ لغت شامل ١,٠٩٥,٩٥٩ کلمه است، همه آن‌ها کلمات عمومی هستند، اما شامل اصطلاحات پزشکی تخصصی نمی‌شوند. در مدل پیشنهادی، متون آموزش، برای ایجاد فرهنگ لغت تخصصی مورد استفاده قرار گرفته است؛ این فرهنگ لغت شامل واژگان تخصصی پیدا شده در سونوگرافی سینه، سونوگرافی سر و گردن و همچنین کلمات تخصصی متون سونوگرافی شکم و لگن است که با ترجمه‌هایی از فرهنگ لغات علوم پرتوشناسی توسط دیوید دُسِت (David J Dowsett) برای شناسایی اشتباهات کلمات تخصصی تکمیل شده است(٤٦). محققان پژوهش حاضر به منظور جلوگیری از تکرار کلمات تخصصی، فرهنگ لغت گسترده‌ی خود را با فرهنگ لغات علوم پرتوشناس با استفاده از برنامه‌ی خاص‌منظور که توسط پژوهشگران این مطالعه تولید شده است، مقایسه نمودند.

به این ترتیب، این اطمینان حاصل شد که هیچ کلمه‌ای بیش از یک‌بار در فرهنگ لغت قرار نگیرد؛ زیرا برخی از کلمات ممکن است در هر دو فرهنگ لغت وجود داشته باشند. بر اساس تجزیه و تحلیل داده‌های آزمون، مشخص گردید که فاصله ویرایش تا ٢ واحد، برای تولید کاندیدا بهینه خواهد بود. با تنظیم فاصله ویرایش به یک، میانگین سه نامزد به عنوان جایگزین‌های بالقوه برای یک خطا تولید شد. با افزایش فاصله ویرایش به ٢، میانگین تعداد نامزدهای تولید شده به ١٥ رسید؛ در نتیجه، زمان محاسبات نیز افزایش یافت. همچنین در پایان، کاندیداهای تولید شده توسط نویسندگان با لغت‌نامه‌ی مرجع، مقایسه شده و در صورت وجود تأیید گردیدند.

مدل کیسه کلمات ممتد (Word2vec CBOW) یک مدل شبکه عصبی یادگیری ماشین سطحی (Shallow) با یک لایه پنهان (Single Hidden Layer) است(٤٧). این مدل برای انجام رتبه‌بندی وابسته به زمینه (Context Dependent) استفاده می‌شود. نحوه‌ی عملکرد مدل در شکل ۲ دیده می‌شود.



شکل ۲: معماری مدل دوگانه جای‌گذاری در بستر مدل کیسه کلمات ممتد

مدل فوق یک کلمه‌ی هدف را در لایه‌ی خروجی از یک زمینه‌ی داده شده در لایه ورودی پیش‌بینی می‌کند. محاسبه لایه پنهان و کلمات هدف شامل دو ماتریس است: ماتریس ورودی و ماتریس خروجی که برای محاسبه‌ی لایه پنهان و کلمات هدف مطابق فرمول ۱ و ۲ استفاده می‌گردند(٤٧).

$$[\mathbf{H}_{1*N}] = [\mathbf{C}_{1*W}] * [\mathbf{IM}_{W*N}] \qquad (۱)$$

$$[\mathbf{T}_{1*W}] = [\mathbf{H}_{1*N}] * [\mathbf{OM}_{N*W}] \qquad (۲)$$

در این جا W نمایانگر تعداد کل کلمات در داده آموزش و N ابعاد لایه پنهان است. در نهایت تابع softmax برای تبدیل لایه خروجی به احتمالات (Pw) برای بروزرسانی و در طول فرایند آموزش با استفاده از روش انتشار-رو-به-عقب (backpropagation) استفاده می شود.، که همچنین به عنوان بردار کلمه شناخته می‌شود، تقریباً در تمامی کاربردهای Word2vec استفاده می‌شود، در حالی که معمولاً نادیده گرفته می‌شود. مطابق فرمول ۳ در روش پیشنهادی ما، هر دو ماتریس و برای محاسبه‌ی میزان قرابت یه کلمه‌ی هدف با زمینه‌های داده شده استفاده می‌گردند.

$$[\mathbf{T}_{1*W}] = [\mathbf{C}_{1*W}] * [\mathbf{IM}_{W*N}] * [\mathbf{OM}_{N*W}] \qquad (۳)$$

به عبارت دیگر، لایه پنهان و کلمات هدف، به ترتیب به عنوان نخستین و دومین جای‌گذاری در نظر گرفته شده‌اند. همچنین تابع softmax استفاده نشد؛ زیرا نیاز به انتشار-رو-به-عقب در این حالت وجود ندارد. در پژوهش جاری مدل کیسه کلمات ممتد بر مبنای جای‌گذاری دوگانه‌ی کلمات، بر روی داده‌های آموزش متون سونوگرافی فارسی آموزش داده شد، تا ماتریس ورودی و ماتریس خروجی را به دست آوریم. برای تولید ماتریس‌های فوق از اندازه پنجره (Window Size) ۷ و اندازه جای‌گذاری (Embedding Size) ۲۰۰ استفاده گردید. بدین‌ترتیب پنجره‌ی محتوا با طول ۳ در هر طرف کلمه‌ی هدف در نظر گرفته شد. در نهایت امتیازات زمینه به دست آمده ممکن است مثبت، صفر، یا منفی باشد. امتیاز زمینه صفر به این معناست که کلمه‌ی هدف، بردار کلمه ندارد.

ماژول تشخیص خطا از دو استراتژی جداگانه بر اساس نوع خطای شناسایی شده استفاده می‌کند. برای خطاهای شکلی، روش جستجوی کلمات در دیکشنری به کار می رود، در حالی‌که خطاهای معنایی از طریق تجزیه و تحلیل زمینه، بررسی می‌شوند. گام اول در تشخیص خطا، بدون توجه به نوع خطا، شامل تشخیص مرز جملات و شناسایی نشانه است. با دریافت جمله‌ی ورودی، مدل ابتدا شروع و پایان جمله را با نشانگرهای شروع جمله، یعنی<آغاز> و

پایان جمله <پایان> به ترتیب علامت‌گذاری می‌کند، سپس تعداد کلمات در جمله را تخمین می‌زند:

(٤) <آغاز> کلمه۱ کلمه۲ ... کلمه پایانی <پایان>

تعداد کلمات، متناظر با حداکثر تعداد چرخه‌هایی است که مدل برای شناسایی یک خطا در جمله اعمال خواهد کرد.

مدل پیشنهادی از روش جستجو در لغت‌نامه برای شناسایی اشتباهات املایی شکلی استفاده می‌کنند. این روش شامل مقایسه‌ی هر کلمه در جمله ورودی با فرهنگ لغت مرجع است که با استفاده از یک جدول هش (Hash Table) ساخته شده است. با شروع از نشانگر <آغاز>، مدل هر کلمه را در جمله برای صحت آن بر اساس دنباله هدف بررسی می‌کند. این فرایند تا زمان رسیدن به نشانگر <پایان> ادامه دارد. در عین حال، اگر یک کلمه به عنوان اشتباه شناسایی شود، فرایند تشخیص خطا متوقف شده و فاز اصلاح خطا آغاز می‌شود. در جدول ۱، نمونه‌ای از تشخیص خطای شکلی آورده شده است:

جدول ۱: تشخیص خطای شکلی توسط مدل پیشنهادی

نمونه‌ای از تشخیص خطای شکلی
در محل بررسی مایغ [مایع] مشاهده نشد.

در مثالی که در جدول ۱ داده شده، کاربر قصد تایپ کلمه «مایع» را داشته اما به اشتباه «مایغ» تایپ شده است. این خطا به دلیل عملیات جایگزینی کاراکتر «غ» با «ع» می‌باشد که در فاصله یک واحدی از کلمه صحیح قرار دارد. مدل پیشنهادی، کلمه «مایغ» را با کلمات موجود در دیکشنری مرجع تطابق داده و با توجه به یافت نشدن در لیست، آن را به عنوان خطا شناسائی شناسایی می‌کند.

در این پژوهش، تجزیه و تحلیل زمینه برای شناسایی خطاهای معنایی مورد استفاده قرار گرفته است. مدل‌های آماری سنتی برای بررسی فرکانس رخداد (frequency) یک کلمه و ارزیابی زمینه کلمه با در نظر گرفتن فرکانس کلمه با «n» عبارت قبلی، به مدل‌های زبان چندگرمی تکیه می‌کردند(٤٨). با این حال، روش‌های جدید از شبکه عصبی برای ارزیابی سازگاری معنایی کلمات در یک جمله‌ی داده شده، استفاده می‌کنند. در روش پیشنهادی، از ویژگی برجسته‌سازی (Highlight) استفاده گردیده و امتیازات زمینه‌ای حاصل از مدل زبان دوجهته، جهت شناسایی و اصلاح خطاهای معنایی مورد بهره قرار گرفته‌اند. فرایند تشخیص خطای معنایی به شرح زیر است:

۱- مدل با تشخیص نشانگر <آغاز> شروع می‌کند و با شروع از کلمه اول جمله، سعی می‌کند هر کلمه را به عنوان یک کلمه‌ی برجسته نشانه‌گذاری کند.

۲- فهرست جایگزین‌های بالقوه برای کلمه‌ی برجسته شده از خروجی مدل دوگانه جای‌گذاری کلمات دریافت می‌شود.

۳- بر اساس سناریوی تولید کاندیدا، نامزدهای جایگزین در فواصل ویرایش ۱ و ۲ از کلمه‌ی برجسته شده تولید می شوند.

٤- فهرست نامزدهای کلمه‌ی هدف، در برابر خروجی مدل پیش‌آموخته برای کلمه‌ی برجسته‌شده مقایسه می‌شود. اگر یکی از نامزدها نسبت به کلمه‌ی برجسته شده احتمال بالاتری داشت، کلمه‌ی اصلی به عنوان یک خطای معنایی تشخیص داده می شود؛ در نتیجه فرآیند تشخیص خطای معنایی به پایان می‌رسد. با این حال، اگر هیچ خطایی شناسایی نشود، مدل یک واحد به سمت چپ حرکت کرده و همان گام‌ها برای تمام کلمات در جمله تکرار می‌شود تا زمانی که با نشانگر <پایان> روبرو شود. بنابراین، لحظه‌ای که خطای معنایی شناسایی شود، فرایند اصلاح فورا آغاز می‌شود؛ سپس، مدل به پردازش جمله بعدی می‌پردازد. در جدول ۲ مثالی از تشخیص خطای معنایی آورده شده است:

جدول ۲: تشخیص خطای معنایی توسط مدل پیشنهادی

نمونه‌ای از تشخیص خطای معنایی
[Highlight]
در سمت چپ توده **اینترارکتال** [اینتراداکتال] دیده شد.

همان‌طور که در جدول ۲ نشان داده شده است، مرحله‌ی اصلاح خطا، زمانی آغاز می‌شود که یک خطا در ورودی شناسایی شود. در این مرحله، از

یک الگوریتم رتبه‌بندی استفاده شده که بر اساس امتیازات زمینه خروجی مدل آموزش‌داده شده عمل می‌کند.

در فرایند اصلاح خطاهای شکلی و معنایی، گام‌های زیر انجام می‌شوند:

۱- ابتدا مدل از فاصله‌ی ویرایش دامارو-لونشتین استفاده می‌کند تا یک مجموعه کاندیدای جایگزین در ۱ یا ۲ واحد فاصله ویرایش را، برای کلمه‌ی اشتباه تولید کند.

۲- سپس کلمه‌ی اشتباه نوشته شده به عنوان Highlight نشانه‌گذاری شده و به عنوان کلمات ماقبل و بعد از آن به عنوان ورودی به مدل آموزش داده شده وارد می‌گردند.

۳- مدل، تمام کلمات احتمالی را از خروجی مدل جای‌گذاری دوگانه کلمات استخراج می کند و آن‌ها را با لیست نامزدها مطابقت می‌دهد.

٤- سپس مدل تعداد معینی از نامزدها با بالاترین امتیازات زمینه را حفظ می‌کند. بر اساس مشاهدات پژوهشگران این مطالعه، تعداد بهینه ۱۰ است.

۵- مدل پیشنهادی، کاندیدای با بالاترین امتیاز تشابه زمینه در خروجی مدل جای‌گذاری دوگانه را به عنوان منتخب با کلمه‌ی اشتباه جایگزین می‌کند.

یافته‌ها

در این بخش، عملکرد روش پیشنهادی را در حوزه‌ی تشخیص و تصحیح خطا ارزیابی کرده و آن را با مدل‌های موجود در حوزه‌ی اصلاح املای کلمات در متون پزشکی زبان فارسی مقایسه می‌گردد. این مقایسه دیدگاه‌های اصلی در خصوص کارآمدی و دقت مدل پیشنهادی در زمینه شناسایی و اصلاح خطاهای املایی در متون سونوگرافی را ارایه خواهد داد.

معیارهای اصلی ارزیابی برای سنجش عملکرد مدل در حوزه‌ی تشخیص خطای شکلی و معنایی و اصلاح آن‌ها، به ترتیب، دقت، بازخوانی، و معیار-F هستند. دقت، میزان دقیق بودن یک مدل را اندازه‌گیری می‌کند؛ در حالی‌که بازخوانی، حساسیت یا جامع بودن آن را برآورد می‌نماید. معیار-F، که میانگین هارمونیک وزن دار هر دو معیار است، با ترکیب آن‌ها محاسبه می شود. در F1، هر دو دقت و بازخوانی وزن برابر دارند. معادله‌ی ۵ معیار-F را توصیف می‌کند.

$$F - Measure = 2 * \frac{P * R}{P + R} \qquad (۵)$$

در این پژوهش، دو مدل پایه برای اصلاح خطاهای شکلی در متون سونوگرافی فارسی پیاده‌سازی شده است تا انجام یک مقایسه جامع میسر شود. این مدل‌ها شامل مدل چهارگرمی (fourgram) یزدانی و همکاران(۳۹)، و یک مدل فارسی کیسه کلمات ممتد است(٤۷). هر دو مدل با استفاده از زبان برنامه‌نویسی پایتون توسعه یافته‌اند و بر روی همان مجموعه داده ای که برای آموزش مدل پیشنهادی استفاده شده است، آموزش داده شده‌اند. هدف مقایسه، بررسی نقاط قوت و ضعف این مدل‌ها و بهره‌برداری از این نتایج برای بهبود فرایند اصلاح خطای املایی در پردازش متون سونوگرافی زبان فارسی است. متأسفانهٔ، برای اصلاح خطای معنایی در حوزهٔ متون پزشکی زبان فارسی تا پیش از این معرفی نشده است؛ ازاین‌رو انجام یک مقایسه جامع در حوزه‌ی تصحیح خطای معنایی دور از دسترس می‌باشد.

در مرحله‌ی نخست از ارزیابی، عملکرد مدل پیشنهادی در زمینه اصلاح خطای شکلی مقایسه گردیده است. چون مدل پیشنهادی برای شناسایی خطاهای شکلی، روش جستجو در لغت‌نامه را به کار می‌برد، معیار-F برای تشخیص خطای املایی ۱۰۰٪ است. جدول ۳ نتایج بررسی عملکرد مدل پیشنهادی را در زمینهٔ اصلاح خطای شکلی بر حسب پارامترهای دقت، بازخوانی و معیار-F ارایه می‌دهد.

جدول ۳: مقایسه‌ی دقت مدل پیشنهادی در زمینه تصحیح خطای شکلی

نام مدل	دقت	بازخوانی	معیار-F
مدل پیشنهادی	٪۹۰/۱	٪۹۱/۵	٪۹۰/۸

بر اساس نتایج جدول ۳ مدل پیشنهادی اوج عملکرد خود را با دقت ۹۰/۱٪ و میزان بازخوانی ۹۱/۵٪ به‌نمایش می‌گذارد. همچنین در این بین مقدار معیار-F برابر با ۹۰/۸٪ می‌باشد که مطلوب است.

همچنین ارزیابی جامع مدل پیشنهادی برای تشخیص و اصلاح خطاهای معنایی در متون سونوگرافی فارسی در جدول ٤ نمایش داده شده است.

جدول ۴: مقایسه‌ی دقت مدل پیشنهادی در زمینه‌ی تصحیح و تشخیص خطای معنایی

نام مدل	نوع وظیفه	دقت	بازخوانی	معیار F-
مدل پیشنهادی	تشخیص خطای معنایی	٪۹۰/۲	٪۹۰/۸	٪۹۰/۵
مدل پیشنهادی	تصحیح خطای معنایی	٪۸۹/۷	٪۹۰/٤	٪۹۰/۰

مدل پیشنهادی بالاترین عملکرد خود را در تشخیص خطای معنایی با کسب معیار-F برابر با ۹۰/۵٪ ارایه داد. همچنین دقت تصحیح خطاهای معنایی مطابق ارزیابی برابر با ۹۰/۰٪ می‌باشد.

پژوهشگران توانایی مدل خود را برای تشخیص و اصلاح خطاهای معنایی ارزیابی نمودند. همان‌طور که در جدول ۲ نشان داده شده، روش پیشنهادی با کسب بالاترین معیار-F برابر با ۹۰/۵٪ و میزان دقت و نرخ بازخوانی برابر با ۹۰/۲٪ و ۹۰/۸٪ در تشخیص خطاهای معنایی، عملکرد مناسبی را نشان داده است. علاوه بر این در بحث تصحیح خطاهای معنایی، میزان دقت و بازخوانی برابر با ۸۹/۷٪ و ۹۰/٤٪ است. همچنین بالاترین امتیاز اصلاح خطاهای معنایی کشف شده توسط مدل پیشنهادی بر اساس معیار-F برابر با ۹۰/۰٪ است که در مقایسه با مدل‌های مورد مقایسه، قابل توجه است.

بحث

اشتباهات املایی از جمله مسایل رایج در گزارش‌های رادیولوژی هستند؛ که اغلب به دلیل وقفه‌های مکرر و شرایط خاص محیط کاری اتفاق می‌افتند. این موارد می‌توانند سلامت بیمار را تهدید کنند، باعث ایجاد ابهام شوند و صحت و اعتبار رادیولوژیست‌ها تأثیر بگذارند(٤۹). هدف اصلی این تحقیق، ابداع یک روش نوآورانه برای تشخیص و اصلاح اشتباهات املایی در متون بالینی فارسی بود. با توجه به ساختار پیچیده و دستور زبان فارسی و نقش حیاتی مستندات دقیق در تضمین ارایه مراقبت‌های مؤثر به بیمار، صحت تحقیق و فراهم نمودن ایمنی بیمار، این امر اهمیت قابل‌توجهی دارد(۲). در بخش تصویربرداری بیمارستان امام خمینی، تولید گزارش‌های رادیولوژی یک فرایند چند مرحله‌ای است که معمولاً حدود ۳۰ دقیقه طول می‌کشد(۵۰). این فرایند، شامل دیکته توسط رادیولوژیست‌ها، تایپ توسط تایپیست‌های پزشکی، و یک فرایند بازبینی و ویرایش قبل از آرشیو نهایی گزارش در سامانه اطلاعات سلامت بیمارستان است. با این حال، این فرایند، شامل فعالیت‌های بدون ارزش افزوده و زمان صرف شده بین تایپ و تأیید نهایی جهت بررسی صحت محتوایی و نیز شکلی متون است که با عنوان مودا (Muda) شناخته می‌شوند(۵۱). مدل پیشنهادی که با دقت قابل توجهی قادر به تشخیص و تصحیح کلمات اشتباه نوشته شده در حین تایپ می‌باشد، به رفع این ناکارآمدی می پردازد و از این طریق زمان بین آماده‌سازی و تأیید نهایی را به شکل قابل توجهی کاهش می‌دهد؛ قابل ذکر است که این امر خود به کمینه کردن مودا نیز منجر می‌شود. روش ارایه شده از یک مدل جای‌گذاری دوگانه‌ی کلمات استفاده می‌کند که به طور خاص برای رفع خطاهای املایی کلمات در متون سونوگرافی تنظیم شده است. این معماری، مدل پیشنهادی را از روش‌های موجود در حوزه‌ی تصحیح متون بالینی فارسی متمایز می‌کند و به آن امکان می‌دهد که به طور مؤثر خطاهای شکلی و معنایی را مدیریت کند.

روش پیشنهادی یزدانی و همکاران از آخرین روش‌های آماری کلاسیک برای اصلاح خطای شکلی است. این مدل به صورت خاص برای اصلاح اشتباهات املایی در متون سونوگرافی فارسی طراحی شده است(۳۹). این روش از یک مدل زبانی چهارگانه‌ی دوجهته‌ی وزن دار برای شناسایی جایگزین مناسب برای یک خطا شکلی بهره می‌برد. مدل فوق مبتنی بر یک معادله‌ی درجه چهار است که اولویت را بر اساس طول دنباله‌ها مشخص می‌کند و منجر به افزایش دقت اصلاح خطای شکلی می‌شود. مطابق ارزیابی انجام شده، میزان پارامترهای دقت، بازخوانی و معیار-F برابر با ۷۵/۷٪، ۷۷/۳٪ و ۷٦/۵٪ می‌باشد.

دیگر مدل مورد ارزیابی، مدل کیسه کلمات ممتد است. این مدل با درک معنای کلمات از طریق تجزیه و تحلیل زمینه آن‌ها کار می‌کند، و سپس این اطلاعات را به عنوان ورودی برای پیش‌بینی کلمات مناسب برای زمینه‌ی داده شده، استفاده می‌کند(٤۷). این مدل طوری طراحی شده است که کلمهٔ هدف (کلمه‌ی مرکز) را بر اساس کلمات زمینه‌ی فراهم شده شناسایی کند. مدل مورد اشاره به طور خاص برای امر تشخیص خطای شکلی آموزش داده شده است. در این مدل از دو ماتریس ورودی و خروجی برای تخمین لایه پنهان استفاده می‌شود. برای آموزش، از داده‌های آموزشی مشابه با مدل پیشنهادی استفاده می‌شود؛ همچنین اندازه‌ی پنجره ۷ و اندازه‌ی ابعاد ۲۰۰ در فاز آموزش مورد

استفاده قرار گرفته‌اند. مطابق ارزیابی انجام گرفته بر روی مدل کیسه کلمات ممتد، مقدار پارامترهای دقت، بازخوانی و همچنین معیار-F به ترتیب برابر با ۸۰/۶٪، ۸۲/۷٪ و ۸۱/۶٪ می‌باشد.

ارزیابی روش پیشنهادی، بیانگر دقت قابل توجه آن در شناسایی و اصلاح خطاهای شکلی در متن بالینی فارسی می‌باشد. از سوی دیگر، در بحث تصحیح خطای شکلی، روش پیشنهادی یزدانی و همکاران کم‌ترین میزان دقت را در بین سه مدل مورد مقایسه نشان می‌دهد که برابر با معیار-F ۷۶/۵٪. می‌باشد؛ بدین‌سان مدل پیشنهادی با ۱۴/۳٪ دقت بیش‌تر، روش فوق را پشت سر می‌گذارد که بار دیگر عملکرد قابل قبول روش آن را تأیید می‌کند. علاوه بر این، روش پیشنهادی ۹/۲٪ دقت بیش‌تر را در مقایسه با مدل کیسه کلمات ممتد زبان فارسی ارایه می‌دهد. این امر مؤید این است که روش دوگانه‌ی جای‌گذاری کلمات به میزان قابل‌توجهی دقیق‌تر از مدل کیسه کلمات ممتد زبان فارسی است.

برای تشخیص خطای معنایی نیز، مدل ارایه‌شده، عملکرد مطلوبی را با به دست آوردن امتیاز معیار-F برابر با ۹۰/۵٪. نشان داد؛ علاوه بر این، مدل به حداکثر امتیاز معیار-F خود ۹۰/۰٪ برای اصلاح این دسته از خطاها رسید. با وجود این نتایج قابل توجه، مدل حاضر محدودیت‌های خاص خود را دارد. همچنین مشخص گردید که در چند مورد، خطاهای کلمه‌ی واقعی تشخیص داده نمی‌شوند؛ این امر در مواقعی اتفاق می‌افتد که خطای معنایی ارتباط زمینه‌ای قوی با دیگر کلمات جمله دارد. به عنوان مثال، در دنباله‌ی کلمات(روده در سمت چپ تستیس چپ دیده شد)، تایپیست پزشکی به اشتباه کلمه‌ی اصلی را(روده) را، به اشتباه با کلمه‌ی(توده) جایگزین کرده است. به این ترتیب خطای معنایی در فاصله‌ی ویرایش ۱ قرار دارد و عبارت تغییر یافته به این شرح می‌باشد:(توده در سمت چپ تستیس چپ دیده شد). از آن‌جایی که این خطای معنایی امتیاز زمینه‌ای بالاتری نسبت به کلمه‌ی اصلی داشته است مدل از در نظر گرفتن آن به عنوان خطای معنایی صرف نظر می‌کند.

برای جلوگیری از نادیده گرفتن چنین خطاهایی، می توان یک لیست از نامزدهای بالقوه همراه با امتیازات زمینه‌ای آن‌ها را به کاربر سیستم ارایه داد که به وی اجازه می‌دهد تا جایگزین مناسب را از میان لیست اولویت‌بندی شده انتخاب کند. این موضوع بر این نکته دلالت دارد که با وجود روند روبه‌رشد مدل‌های پیشرفته، وجود و بهره‌جویی از تخصص انسانی در اغلب موارد ضروری است.

علاوه بر این، با آن‌که مدل معرفی شده در این پژوهش، در تشخیص و تصحیح خطاهای شکلی و معنایی عملکرد مؤثری دارد، اما قابلیت مدیریت خطاهای دستور زبانی را ندارد. همچنین، مدل پیشنهادی مانند اغلب مدل‌های این حوزه قادر به تشخیص و تصحیح تنها یک خطا در هر جمله است و نمی‌تواند بیش از یک خطا در هر جمله را مدیریت کند. همچنین طی بررسی‌ها، مواردی رویت شد که در آن‌ها در یک جمله، دو و یا بیش‌تر خطا وجود داشت. برای رفع این محدودیت‌ها، پژوهش‌های آینده می‌توانند چگونگی ادغام ویژگی‌های اختصاصی حوزه‌ی تصحیح خطا در مدل‌های خود بررسی کنند تا عملکرد آن را بهبود بخشند.

در حالی‌که مدل‌های اصلاح املای فعلی مختص متون عمومی زبان فارسی هستند و خاص حوزه‌ی پزشکی نیستند، مدل پیشنهادی به طور خاص بر تشخیص و تصحیح خودکار اشتباهات املایی در گزارش‌های سونوگرافی فارسی تمرکز یافته است. به باور نویسندگان، ادغام سامانه‌های املای خودکار در سامانه‌های حیاتی مدیریت سلامت بیماران مانند بیماری‌های آلرژیک، نسخه‌های دارو، مباحث تشخیصی و... می‌تواند به طور قابل توجهی کیفیت و دقت سوابق پزشکی الکترونیک را بهبود بخشد. همچنین مدل پیشنهادی می‌تواند به عنوان یک برنامه‌ی افزودنی در Microsoft Office Word و انواع مروگرهای اینترنتی نصب شود و از طریق API در سامانه‌های اطلاعات سلامت بیمارستان‌ها استفاده گردد. این کاربردهای بالقوه، امکان استفاده و تعمیم مدل پیشنهادی را حتی فراتر از زمینه متون بالینی فراهم می‌نماید.

نتیجه‌گیری

در این پژوهش، روشی جدید برای تشخیص و اصلاح هر دو خطای املایی شکلی و معنایی در متون سونوگرافی زبان فارسی ارایه گردید. مدل ارایه‌شده از یک مدل پیشرفته‌ی جای‌گذاری دوگانه‌ی کلمات استفاده می‌کند که به طور خاص برای امر اصلاح املایی در حوزه‌ی متون سونوگرافی زبان فارسی تنظیم شده است. مطابق ارزیابی انجام‌گرفته، روش پیشنهادی در شناسایی و اصلاح خطاهای املایی و معنایی در متون سونوگرافی فارسی، دقت قابل قبولی را ارایه می‌کند که این امر موید پتانسیل کاربرد عملی آن در این حوزه‌ی خاص و حساس می‌باشد.

در آینده نویسندگان مقاله بر آن هستند تا به بحث تشخیص و تصحیح خودکار

چندین خطا در یک جمله بپردازند و استراتژی‌هایی را برای مدیریت سایر خطاها، از جمله خطاهای دستور زبان، ارایه نمایند. همچنین استفاده از اطلاعات آوایی و فاصله کاراکترها بر روی صفحه‌کلید نیز می‌توانند در پژوهش‌های آینده مورد بررسی قرار گرفته تا دقت تصحیح خطاهای متنی در متون بالینی فارسی را بیش از پیش بهبود بخشند: با رفع این چالش‌ها و بهبود دقت روش پیشنهادی، مدلی جامع و مناسب برای مدیریت خطاهای املایی در کلیه حوزه‌های پزشکی در دسترس خواهد بود.

تشکر و قدردانی

بدین‌وسیله پژوهشگران بر خود لازم می‌دانند از تمامی کسانی که در انجام هر چه بهتر این پژوهش همکاری نمودند، تشکر و قدردانی نمایند. پژوهش حاضر بخشی از پایان‌نامه با عنوان «روشی جدید جهت تصحیح خودکار غلط‌های متنی با استفاده از مدل‌سازی زبانی و اطلاعات معنایی» در مقطع دکتری رشته مهندسی کامپیوتر-گرایش نرم افزار، بخش مهندسی کامپیوتر دانشگاه آزاد اسلامی واحد کرمان مصوب با کد اخلاق IR.IAU.KERMAN.REC.1402.124 است.

References

1. Hossain E, Rana R, Higgins N, Soar J, Barua PD, Pisani AR, et al. Natural language processing in electronic health records in relation to healthcare decision-making: A systematic review. Computers in Biology and Medicine 2023; 155(1): 106649.

2. Sanderson AL & Burns JP. Clinical documentation for intensivists: The impact of diagnosis documentation. Critical Care Medicine 2020; 48(4): 579-87.

3. Bravo-Candel D, Lopez-Hernandez J, Garcia-Diaz JA, Molina-Molina F & Garcia-Sanchez F. Automatic correction of real-word errors in Spanish clinical texts. Sensors (Basel, Switzerland) 2021; 21(9): 2893.

4. Dashti SMS, Khatibi-Bardsiri A & Jafari-Shahbazzadeh M. PERCORE: A deep learning-based framework for persian spelling correction with phonetic analysis. International Journal of Computational Intelligence Systems 2024; 17(1): 114.

5. Dashti SMS, Fakhrahmad SM, Sadreddini MH & Golkar A. Toward a thesis in automatic context-sensitive spelling correction. International Journal of Artificial Intelligence and Mechatronics 2014; 3(1): 19-24.

6. Dashti SMS, Khatibi-Bardsiri A & Khatibi-Bardsiri V. Correcting real-word spelling errors: A new hybrid approach. Digital Scholarship in the Humanities 2018; 33(3): 488-99.

7. Dashti SMS. Real-word error correction with trigrams: Correcting multiple errors in a sentence. Language Resources and Evaluation 2018; 52(1): 485-502.

8. Pande H. Effective search space reduction for spell correction using character neural embeddings. Valencia, Spain: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017.

9. Jayanthi SM, Pruthi D & Neubig G. NeuSpell: A neural spelling correction toolkit. Available at: https://arxiv.org/abs/2010.11085. 2020.

10. Lee JH, Kim M & Kwon HC. Deep learning-based context-sensitive spelling typing error correction. IEEE Access 2020; 8(1): 152565-78.

11. Liu S, Yang T, Yue T, Feng Z & Wang D. PLOME: Pre-training with misspelled knowledge for Chinese spelling correction. China: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021.

12. Sun R, Wu X & Wu Y. An error-guided correction model for chinese spelling error correction. Abu Dhabi, United Arab Emirates: Conference Findings of the Association for Computational Linguistics (EMNLP), 2022.

13. Wang X, Liu Y, Li J, Miljanic V, Zhao S & Khalil H. Towards contextual spelling correction for customization of end-to-end speech recognition systems. IEEE/ACM Transactions on Audio, Speech, and Language Processing 2022; 30(1): 3089-97

14. Zhang R, Pang C, Zhang C, Wang S, He Z, Sun Y, et al. Correcting Chinese spelling errors with phonetic pre-training. Bangkok, Thailand: Conference Findings of the Association for Computational Linguistics (ACL-IJCNLP), 2021.

15. Li J, Wu G, Yin D, Wang H & Wang Y. MDCSpell: A multi-task detector-corrector framework for Chinese spelling correction. Dublin, Ireland: SIGIR '21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (ACL), 2021.

16. Mridha MF, Lima AA, Nur K, Das SC, Hasan M & Kabir MM. A survey of automatic text summarization: Progress, process and challenges. IEEE Access 2021; 9(1): 156043-70.

17. Chen YP, Chen YY, Lin JJ, Huang CH & Lai F. Modified bidirectional encoder representations from transformers extractive summarization model for hospital information systems based on character-level tokens (AlphaBERT): Development and performance evaluation. JMIR Medical Informatics 2020; 8(4): e17787.

18. Dalianis H. Clinical text mining: Secondary use of electronic patient records. Switzerland: Springer Nature; 2018: 10-4.

19. Ruch P, Baud R & Geissbuhler A. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. Artificial Intelligence in Medicine 2003; 29(1-2): 169-84.

20. Siklosi B, Novak A & Proszeky G. Context-aware correction of spelling errors in Hungarian medical documents. Computer Speech and Language 2016; 35(1): 219-33.

21. Grigonyte G, Kvist M, Velupillai S & Wiren M. Improving readability of Swedish electronic health records through lexical simplification: First results. Gothenburg, Sweden: Conference Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR), 2014.

22. Zhou X, Zheng A, Yin J, Chen R, Zhao X, Xu W, et al. Context-sensitive spelling correction of consumer-generated content on health care. JMIR Medical Informatics 2015 ; 3(3): e27.

23. Tolentino HD, Matters MD, Walop W, Law B, Tong W, Liu F, et al. A UMLS-based spell checker for natural language processing in vaccine safety. BMC Medical Informatics and Decision Making 2007; 7(1): 1-3.

24. Wong W & Glance D. Statistical semantic and clinician confidence analysis for correcting abbreviations and spelling errors in clinical progress notes. Artificial Intelligence in Medicine 2011; 53(3): 171-80.

25. Doan S, Bastarache L, Klimkowski S, Denny JC & Xu H. Integrating existing natural language processing tools for medication extraction from discharge summaries. Journal of the American Medical Informatics Association 2010; 17(5): 528-31.

26. Lai KH, Topaz M, Goss FR & Zhou L. Automated misspelling detection and correction in clinical free-text records. Journal of Biomedical Informatics 2015; 55(1): 188-95.

27. Fivez P, Suster S & Daelemans W. Unsupervised context-sensitive spelling correction of clinical free-text with word and character n-gram embeddings. Vancouver, Canada: BioNLP, 2017.

28. D-Hondt E, Grouin C & Grau B. Low-resource OCR error detection and correction in French clinical Texts. Austin, Texas, United States of America: Conference Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis, 2016.

29. Pérez A, Atutxa A, Casillas A, Gojenola K & Sellart Á. Inferred joint multigram models for medical term normalization according to ICD. International journal of medical informatics 2018; 110(1): 111-17.

30. Kashefi O, Sharifi M & Minaie B. A novel string distance metric for ranking Persian respelling suggestions. Natural Language Engineering 2013;19(2): 259-84.

31. Faili H, Ehsan N, Montazery M & Pilehvar MT. Vafa spell-checker for detecting spelling, grammatical, and real-word errors of Persian language. Digital Scholarship in the Humanities 2016; 31(1): 95-117.

32. Dastgheib MB, Fakhrahmad SM & Zolghadri-Jahromi M. Perspell: A new Persian semantic-based spelling correction system. Digital Scholarship in the Humanities 2017; 32(3): 543-53.

33. Ghayoomi M & Assi SM. Word prediction in a running text: A statistical language modeling for the Persian language. Sydney, Australia: Proceedings the Australasian Language Technology Workshop, 2005.

34. Ghayoomi M, Momtazi S & Bijankhan M. A study of corpus development for Persian. International Journal on Asian Language Processing 2010; 20(1): 17-34.

35. Mosavi-Miangah T. FarsiSpell: A spell-checking system for Persian using a large monolingual corpus. Literary and Linguistic Computing 2014; 29(1): 56-73.

36. Naseem T & Hussain S. A novel approach for ranking spelling error corrections for Urdu. Language Resources and Evaluation 2007; 41(2): 117-28.

37. Shamsfard M. Challenges and open problems in Persian text processing. Poznan, Poland: Proceedings of the 5th Language and Technology (LTC), 2011.

38. Shamsfard M, Jafari HS & Ilbeygi M. STeP-1: A set of fundamental tools for Persian text processing. Valletta, Malta: Proceedings of the International Conference on Language Resources and Evaluation (LREC), 2010.

39. Yazdani A, Ghazisaeedi M, Ahmadinejad N, Giti M, Amjadi H & Nahvijou A. Automated misspelling detection and correction in Persian clinical text. Journal of Digital Imaging 2020; 33(3): 555-62.

40. Yazdani A, Safdari R, Golkar A & Rostam-Niakan-Kalhori S. Words prediction based on N-gram model for free-text entry in electronic health records. Health Information Science and Systems 2019; 7(1): 1-7.

41. Dehkhoda AA. Dehkhoda dictionary. Tehran: Tehran University. Available at: https://icps.ut.ac.ir/fa/dictionary. 1993.

42. Damerau FJ. A technique for computer detection and correction of spelling errors. Communications of the ACM 1964; 7(3): 171-6.

43. Peterson JL. A note on undetected typing errors. Communications of the ACM 1986; 29(7): 633-7.

44. Huang Y, Murphey YL & Ge Y. Automotive diagnosis typo correction using domain knowledge and machine learning. Singapore: IEEE Symposium on Computational Intelligence and Data Mining (CIDM), 2013.

45. Kukich K. Techniques for automatically correcting words in text. ACM Computing Surveys (CSUR) 1992; 24(4): 377-439.

46. Dowsett D. Radiological sciences dictionary: Keywords, names and definitions. London: CRC Press; 2009: 29-350.

47. Mikolov T, Chen K, Corrado G & Dean J. Efficient estimation of word representations in vector space. Available at: https://arxiv.org/abs/1301.3781. 2013.

48. Oralbekova D, Mamyrbayev O, Othman M, Kassymova D & Mukhsina K. Contemporary approaches in evolving language models. Applied Sciences 2023; 13(23): 12901.

49. Minn MJ, Zandieh AR & Filice RW. Improving radiology report quality by rapidly notifying radiologist of report errors. Journal of Digital Imaging 2015; 28(4): 492-8.

50. Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY & Lungren MP. CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. Available at: https://arxiv.org/abs/2004.09167. 2020.

51. Kruskal JB, Reedy A, Pascal L, Rosen MP & Boiselle PM. Quality initiatives: lean approach to improving performance and efficiency in a radiology department. Radiographics 2012; 32(2): 573-87.

Journal of Payavard Salamat
Vol. 18, No. 1, Apr 2024

Automatic Spelling Correction in Persian Sonography Text with Neural Networks

Seyed Mohammad Sadegh Dashti[1] (Ph.D.), Amid Khatibi Bardsiri[2*] (Ph.D.), Mehdi Jafari Shahbazzadeh[3] (Ph.D.)

1 Ph.D. in Computer Engineering, Kerman Branch, Islamic Azad University, Kerman, Iran
2 Assistant Professor in Computer Engineering, Department of Computer Engineering, Kerman Branch, Islamic Azad University, Kerman, Iran
3 Assistant Professor in Electrical Engineering, Department of Electrical Engineering, Kerman Branch, Islamic Azad University, Kerman, Iran

Abstract

Received: 9 Oct. 2023
Accepted: 15 Mar. 2024

Background and Aim: Medical reports and electronic health records are critically important for diagnosis, treatment, patient protection, and medical research. Correcting spelling errors in medical texts is essential to ensure accurate interpretation of information. This research was conducted to automatically correct spelling mistakes in Persian medical texts using neural networks.

Material and Methods: In this study, which was conducted in 2023, a computational model based on artificial intelligence neural networks and dual embedding techniques was developed using Python in a Windows environment. The dual embedding model was fine-tuned for correcting spelling errors in Persian sonography texts. The proposed model employs various techniques for automatic error detection, including dictionary lookup approach and contextual similarity coefficients. Furthermore, features specific to text processing, such as Edit-Distance, along with similarity coefficients, were utilized to automatically select the most appropriate substitute for a misspelled word. The training and testing data for the current model were sourced from a collection of sonography texts from the Imam Khomeini Hospital's sonography clinic in Tehran.

Results: The proposed model which is based on artificial neural networks, leverages a novel dual-embedding architecture to select the best candidate words for correcting both non-word and real-word errors. According to the evaluation results on Persian sonography text, the proposed model achieved an F-Measure accuracy of 90.5% in detecting real-word errors. Furthermore, it demonstrated an impressive 90% accuracy in automatically correcting these real-word errors. Additionally, the model exhibited a strong performance, achieving 90.8% accuracy in correcting non-word errors.

Conclusion: Based on the evaluation results, the proposed method is robust against various changes in word forms and can effectively manage a wide range of morphological and semantic errors, including replacements, transpositions, insertions, and deletions in medical texts. The integration of Edit-Distance with textual similarity coefficients extracted from the dual embedding model significantly enhanced the accuracy of spelling corrections in Persian sonography texts, ensuring greater validity of such documents. The authors believe that the proposed model represents a significant advancement in the detection and correction of spelling errors in Persian sonography texts.

Keywords: Spelling Correction, Neural Embeddings, Neural Networks, Sonography Text, Persian Language Processing

*** Corresponding Author:**
Khatibi Bardsiri A
Email:
a.khatibi@srbiau.ac.ir