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

پیاده‌سازی یک سیستم کنترل خطای املایی در زبان فارسی بر اساس کدگذاری Soundex

انور بهرام‌پور[۱]، فردین اخلاقیان طاب[۲]، جلال سجادی[۲]، وفا بارخدا[۲]
[۱]گروه فناوری اطلاعات، دانشگاه آزاد اسلامی- واحد سنندج، سنندج، ایران Bahrampour@iausdj.ac.ir
[۲]گروه کامپیوتر و فناوری اطلاعات، دانشگاه کردستان، سنندج، ایران{Akhlaghian,Jsajadi,Barkhoda@uok.ac.ir}

چکیده

در بسیاری از کاربردها مانند پردازشگرهای متن، موتورهای جستجو، لغتنامه‌های الکترونیك، کاربردهای تلفن همراه و ... عمل کنترل املا بخشی از عملکرد کلی سیستم بشمار می‌رود. از آنجا که کارایی الگوریتم‌های موجود در زبانهای مختلف متفاوت بوده و این راهکارها نیازمند تعریف ساختارها و الگوریتمهای متناسب با زبان است، بررسی ساختارها و الگوریتمهای کنترل املا در زبان فارسی نیز بسیار ضروری است. در این مقاله انواع خطاهای املایی، ساختار سیستمهای کنترل املا و تعدادی از الگوریتمهای مورد استفاده در تشخیص و تصحیح خطاهای املایی بررسی شده است. سپس با تعریف یك سیستم کدگذاری بر اساس روش Soundex برای زبان فارسی، یك سیستم کنترل املا در زبان فارسی با کارایی قابل قبول پیاده سازی شده است. سیستم پیشنهادی برای ذخیره کردن لغات در فرهنگ‌لغت از درخت B استفاده می‌نماید.

کلمات کلیدی

سیستم کنترل خطا، تشخیص خطا، تصحیح خطا، روش کدگذاری Soundex

۱. مقدمه

سیستم‌های کنترل خطا[1] به برنامه‌هایی گفته می‌شوند که کلماتی را که امکان دارد از لحاظ املایی غلط باشند، علامت‌گذاری می‌کنند و به ازای هر یك از کلمات نادرست تعدادی واژه مشابه به عنوان جایگزین پیشنهاد می‌دهند. عملکرد این برنامه‌ها شامل دو بخش تشخیص خطا و تصحیح خطا است[1]. این برنامه‌ها ممکن است به عنوان نرم‌افزاری مستقل کار تشخیص و تصحیح خطا را انجام دهند و یا اینکه بخشی از برنامه‌های بزرگتر همانند پردازشگرهای متن، لغتنامه‌های الکترونیك و یا موتورهای جستجو باشند. سیستم‌های کنترل املا برای یافتن و تصحیح خودکار خطاهای املایی در متن زبان‌های طبیعی، بسیار مورد ملاحظه هستند و به طور پیوسته بر محبوبیت آنها افزوده می‌شود. به این برنامه‌ها با عنوان تصحیح‌کننده خودکار خطا[2] نیز ارجاع می‌شود[2]. تلاش برای یافتن الگوریتمها و تکنیکهایی برای تشخیص و تصحیح خودکار خطاهای املایی با گسترش ویرایشگرهای متنی، در اوایل دهه ۱۹۶۰ شروع شد[1] و به تدریج پس از ظهور سیستمهای با قابلیت محدود تشخیص خطا، سیستمهای با امکان تشخیص و تصحیح خطا و قابل نصب بر روی ویرایشگرهای مختلف و نیز قابلیت تصحیح جملات از نظر گرامری پا به عرصه وجود گذاشتند.

یک سیستم کنترل املای ساده هر واژه را در محتوای فرهنگ لغت جستجو می‌کند، اگر آن را در فرهنگ لغت پیدا نمود آن واژه و یا ریشه آن یک کلمه معتبر بوده در غیر این صورت آنرا بعنوان یک خطای املایی شناسایی می‌کند؛ بدیهی است که تمامی واژه‌های زبان در فرهنگ لغت وجود ندارد، برای مثال ریشه افعال در فرهنگ لغت قرار گرفته و شکل‌های مختلف آنرا باید با اعمال قواعد زبان تولید نمود، بنابراین عملکرد یک سیستم کنترل املا فقط یک جستجوی ساده در فرهنگ لغت نیست [1,2,3] البته عمل تشخیص خطای املایی محدود به استفاده از فرهنگ لغت نیست و روشهای دیگری نیز برای تشخیص خطا وجود دارد که در بخشهای بعدی به تعدادی از آنها اشاره شده است. با شناسایی واژه نادرست در یک متن تلاش برای یافتن واژه صحیح در سیستم کنترل املا شروع می‌گردد. اگر کلمه‌ای در لغتنامه یافت نشد، آن را به عنوان کاندیدای خطا در نظر گرفته و در این صورت امکان دارد عملکرد سیستم اینگونه باشد که لیستی از کلمات مشابه و نزدیك به آن کلمه را به ترتیب میزان مشابهت برای کاربر تولید کند. برای تعیین میزان مشابهت کلمات از الگوریتم‌های مختلفی استفاده می‌شود که در این میان

[1] Spell Checker

[2] Automatic Error Corrector