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

تبدیل متن فارسی به زنجیره واجی با استفاده از تحلیلگر صرفی

وحید مواجی[۱]، محرم اسلامی[۲]،
[۱]مرکز زبان‌ها و زبان‌شناسی، دانشگاه صنعتی شریف mavaij@alum.sharif.edu
[۲] دانشگاه زنجان melsami@znu.ac.ir

چکیده

در مقاله حاضر می‌کوشیم روشی خودکار برای تبدیل متون فارسی به زنجیره واجی ارائه دهیم. خط فارسی به دلیل دشواری‌های پردازشی که دارد ورودی مناسبی برای برنامه‌های پردازش متن به حساب نمی‌آید. از ویژگی‌های خط فارسی می‌توان به عدم نمایش واکه‌های کوتاه و به دنبال آن موضوع هم‌نویسه‌گی، مسأله کسره اضافه، فاصله بینِ اجزای کلمه واحد، فقدان فاصله بین کلمه‌های مستقل، موضوعِ جدانویسی و پیوسته‌نویسی و غیره اشاره کرد. برخورداری خط فارسی از ویژگی‌های که برشمردیم موجب می‌شود قبل از انجام هرگونه پردازشی، متون فارسی را به زنجیره واجی تبدیل کنیم. خروجی برنامه تبدیل متن به زنجیره واجی کاربردهای متعددی منجمله در تبدیل خودکار متن به گفتار، واج‌نویسی صحیح متون، آموزش زبان فارسی به غیرفارسی‌زبانان، فرهنگ نویسی و غیره دارد. در این مقاله با استفاده از تحلیلگر صرفی پارس-مورف که توسط نگارندگان طراحی و پیاده‌سازی شده است، متن ورودی از لحاظ صرفی تحلیل شده و و اجزای صرفی آن از قبیل پیشوندها، پسوندها، اشتقاق و ترکیب بدست آمده و سپس با استفاده از واژگان زایای زبان فارسی، صورت واجی آنها با هم ترکیب شده و در نهایت صورت واجی متن ورودی به‌دست می‌آید.

کلمات کلیدی

متن فارسی، زنجیره واجی، تحلیلگر صرفی، تبدیل متن به گفتار

۱- مقدمه

یک سامانه تبدیل متن به گفتار از دو قسمت تبدیل متن به زنجیره واج‌های تشکیل‌دهنده آن و نیز قسمت تبدیل زنجیره واج‌ها به گفتار تشکیل می‌گردد. در این مقاله روی قسمت اول یعنی تبدیل متن به زنجیره واجی تمرکز داریم. روش‌های متعددی برای تبدیل متن به زنجیره واجی مورد استفاده قرار گرفته است. در [1] از قواعد تبدیل نویسه به صورت واجی استفاده شده و استثنائات نیز از یک فرهنگ استخراج می‌شود. استفاده از یک درخت تصمیم چندسطحی که هر نویسه را نسبت به حروف مجاور آن به صورت یک درخت نمایش می‌دهد در [2] مورد مطالعه قرار گرفته است. استفاده از روش‌های زبان طبیعی نیز مورد بررسی قرار گرفته است. در این روش، هر تکواژ به همراه اطلاعات مربوط به صورت‌های صرفی مختلف آن مانند صورت جمع، گذشته، حال، و غیره و کلیه اطلاعات صرفی مربوطه در یک دادگان ذخیره می‌گردد. در این حالت نیز اگر کلمه در فهرست تکواژها موجود نبود از قواعد نویسه به صورت واجی یا فرهنگ استثناها استفاده می‌شود [۳].

در تمامی تحقیقات مربوط به پردازش‌های خودکار زبانی در زبان فارسی، به خصوص در پردازش متن فارسی برای مقاصد مختلف اعم از ترجمه ماشینی، تبدیل متن به گفتار و غیره از نوعی تحلیل‌گر صرفی استفاده می‌شود؛ اگر چه در اغلب مواقع تحلیل‌گرهای صرفی در تحقیقات پیشین محدود، هدف‌محور و فاقد پشتوانه جامع زبان‌شناختی است. به عنوان مرور پیشینه پژوهش در ادامه تنها به مواردی اشاره می‌کنیم که در تحلیل ساخت درونی کلمه فارسی نگاه ساخت‌مند داشته‌اند و با یک رویکرد زبان‌شناختی- مهندسی به تحلیل صرفی کلمه فارسی پرداخته‌اند. در این خصوص ابتدا می‌توان به مطالعات دقیقی در پروژه شیراز [4] در طراحی سامانه ترجمه ماشینی فارسی-انگلیسی اشاره کرد که در میانه راه متوقف شد. دومین مورد از طراحی تحلیل خودکار فارسی مربوط به تحلیل‌گر تصریفی زبان فارسی است که در سامانه تبدیل متن به گفتار فارسی "گویا" به کار گرفته شد [۵] که تنها به تحلیل تصریفی کلمه محدود می‌شد. آماده‌سازی متن معیار برای زبان فارسی ۱ با عنوان اختصاری STeP1 سامانه دیگری است که قادر است کلمات فارسی را از نظر صرفی تجزیه و تحلیل کند [۶]. STeP1 با استفاده از واژگان زایای زبان فارسی [۷] و قواعد صرفی که طراحان آن در نظر گرفته‌اند کار می‌کند.

در این مقاله روشی جدید برای تبدیل خودکار متن فارسی به زنجیره فارسی ارائه می‌شود. در این روش با استفاده از تحلیلگر صرفی که طراحی کرده‌ایم عبارت ورودی را به اجزای