تبدیل متن فارسی به زنجیره واجی با استفاده از تحلیلگر صرفی

وحید مواجی[۱] و محرم اسلامی[۲]

[۱]دانشگاه صنعتی شریف
[۲]دانشگاه زنجان

۱۵ شهریور ۱۳۹۱

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.

کلیدواژه‌ها: 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 Allen (1992) 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 Torkkola (1993) 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 (صیادیان و نصیرزاده، ۱۳۷۵).

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 Megerdoomian (2000) 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 (اسلامی و دیگران، ۱۳۸۳b) که تنها به تحلیل تصریفی کلمه محدود می‌شد. آماده‌سازی متن معیار برای زبان فارسی ۱ با عنوان اختصاری STeP۱ سامانه دیگری است که قادر است کلمات فارسی را از نظر صرفی تجزیه و تحلیل کند (شمس‌فرد، ۱۳۸۸). STeP۱ با استفاده از واژگان زایای زبان فارسی (اسلامی و دیگران، ۱۳۸۳a) 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 تشکیل‌دهنده آن تقطیع می‌کنیم و سپس هر قطعه را به عنوان ورودی به تحلیلگر صرفی می‌دهیم تا به اجزای صرفی تشکیل‌دهنده خود از قبیل پسوندها و پیشوندهای تصریفی و اشتقاقی، پایه‌های ترکیب و اشتقاق و غیره تجزیه شود و سپس صورت واجی اجزا با هم ترکیب شده و زنجیره واجی کل متن به دست می‌آید.

۲ تحلیلگر صرفی

در این تحقیق پس از صورت‌بندی ساخت درونی کلمه، سامانه‌ای را طراحی کرده‌ایم که قادر است به طور خودکار ساخت درونی کلمات فارسی را تحلیل و اجزای سازنده کلمه را مشخص کند. در طراحی پارس مورف، ساخت تصریفی کلمه در زبان فارسی (اسلامی و علیزاده‌لمجیری، ۱۳۸۸) به صورت ساختمند در نظر گرفته شده است و جایگاه طبقات مختلف وندهای تصریفی در ساختمان انواع کلمات مشخص شده است. پارس مورف در تحلیل تصریفی کلمه ابتدا ستاک را شناسایی می‌کند و متناسب با ساخت تصریفی پیش‌بینی شده برای آن نوع ستاک به دنبال انواع وندهای تصریفی در جایگاه‌های خاص، با در نظر گرفتن صورت‌های مختلف نوشتاری هرکدام از طبقات، می‌گردد. یافتن وندی مثلاً در جایگاه دوم پس از ستاک اسمی نشان می‌دهد که جایگاه اول خالی مانده است.

اگر ستاک پیچیده باشد، پارس مورف با اعمال قواعد واژه‌سازی زبان فارسی، ستاک مورد نظر را از حیث مشتق یا مرکب بودن تجزیه و تحلیل و اجزای سازنده آن را با ذکر مقوله دستوری و نقش آنها مشخص می‌کند. به دلیل معتبر نبودن فاصله به عنوان مرز کلمه در متون فارسی، پارس مورف در تجزیه ستاک‌های پیچیده، ترکیب‌های بالقوه را نیز به عنوان گزینه‌های بعدی در اختیار ما می‌گذارد. مثلاً ستاک پیچیده "کارگر" در بخش اشتقاق به عنوان کلمه مشتق شناسایی می‌شود و یا در ترکیب گفته می‌شود که در فارسی ممکن است آن ترکیب "کار (N۱) + گر (Adv)" یعنی اسم + قید باشد که این دو کلمه بی‌فاصله در کنار هم آمده‌اند.

پارس مورف، تحلیل‌گر صرفی زبان ،فارسی بر یک مبنای کاملاً علمی زبانی استوار است و در چارچوب ساختمان صرفی کلمه فارسی به تجزیه و تحلیل و تعیین نقش هر کدام از اجزا در درون کلمه می‌پردازد. پس از صورت‌بندی دقیق اطلاعات نظام تصریف و واژه‌سازی در زبان فارسی سعی کردیم در مرحله اجرا و طراحی سامانه پارس مورف تمامی آن اطلاعات را به شکل دقیق به کار بگیریم. در حال حاضر پارس مورف با استفاده از آخرین ویرایش واژگان زایای زبان فارسی (اسلامی و دیگران، ۱۳۸۳a) که حدود ۴۵۰۰۰ واژه در آن قرار دارد و در چارچوب قواعد صرفی فارسی که در اختیار دارد، با دقت بالای ۹۵٪ می‌تواند ساخت درونی کلمات فارسی را تحلیل کند. نیز می‌تواند با استفاده از امکانات و اطلاعات صرفی که در اختیار دارد کلمات خارج از واژگان را نیز از حیث تصریف و واژه‌سازی تجزیه و تحلیل کند.

۳ تبدیل متن به زنجیره واجی

در برنامه تحلیلگر صرفی پارس-مورف گزینه‌ای به نام Phonology وجود دارد که از طریق آن متن فارسی به عنوان ورودی برنامه داده شده و نتیجه به صورت زنجیره واجی نشان داده می‌شود. برای علائم واجی زبان فارسی از علائم موجود در (ثمره، ۱۳۷۸) استفاده کرده‌ایم. مثلاً واج‌های ʔ، š، â، ǰ، č و ž به ترتیب نمایانگر نویسه‌های /همزه/، /ش/، /آ/، /ج/، /چ/ و /ژ/ می‌باشند.

خروجی برنامه برای جمله "آن یکی نحوی به کشتی درنشست" در جدول ۱ آمده است. همانطور که مشاهده می‌گردد، دو صورت واجی برای این عبارت نشان داده شده است که در یکی کلمه "کشتی" به صورت / kašti / و در دیگری به صورت / kešti / واج‌نویسی شده است.

ʔân yeki nahvi beh kašti dar nešast
ʔân yeki nahvi beh kešti dar nešast

جدول ۱: زنجیره واجی عبارت "آن یکی نحوی به کشتی درنشست"

در جدول ۲، زنجیره واجی عبارت "مردم حضور دارند" آمده است. در این حالت هم دو نوع واج‌نویسی برای کلمه "مردم" آمده است: اولی دارای مقوله دستوری اسم است که به صورت /mardom/ نوشته شده و دیگری نتیجه تحلیل صرفی به صورت فعل "مردن" در حالت اول شخص مفرد است یعنی .mord+am از آنجا که تحلیل نحوی روی عبارت ورودی انجام نمی‌شود، این زنجیره واجی نیز از لحاظ برنامه معتبر است. ولی با افزودن اطلاعات نحوی به برنامه می‌توان دقت آن را بالاتر برد.

mardom hozur dârand
mordam hozur dârand

جدول ۲: زنجیره واجی عبارت "مردم حضور دارند"

مسأله بعدی، مسأله کسره اضافه است که باید تمهیدی برای آن اندیشیده شود. چون کسره اضافه در متن فارسی نمایش داده نمی‌شود، ولی در زنجیره واجی حضور دارد؛ لازم است اطلاعات نحوی به برنامه تبدیل متن به زنجیره واجی اضافه شود تا دقت برنامه افزایش یابد. برای مثال در جدول ۳ خروجی برنامه به ازای عبارت "کتاب من کو؟" آورده شده که در آن کسره اضافه را نمی‌بینیم.

ketâb man ku

جدول ۳: زنجیره واجی عبارت "کتاب من کو؟"

۴ نتیجه

در این مقاله، با استفاده از سامانه تحلیل‌گر صرفی کلمه در زبان فارسی با عنوان پارس مورف، که بر پایه یک مطالعه دقیق زبان‌شناختی از نظام صرفی زبان فارسی استوار است، سعی کردیم تا فرایند خودکار تبدیل متن فارسی به زنجیره واجی را توسعه دهیم. پارس مورف قادر است ساخت درونی کلمه فارسی را از حیث نظام تصریف و واژه‌سازی تجزیه و تحلیل کند و برای هر کدام از اجزا در درون کلمه عنوان و نقش زبانی خاصی اختصاص دهد.

از چالش‌های پیش رو که در کارهای آتی باید به آنها پرداخته شود، درنظر گرفتن دشواری‌های موجود در پردازش متن فارسی (اسلامی، ۱۳۸۱) است که باید راهکارهایی برای حل آن اندیشیده شود. از نتایج این کار می‌توان در زمینه انواع پردازش خودکار زبان فارسی و به طور مشخص در تبدیل متن به گفتار استفاده نمود.

مراجع

Allen, J. (1992), "Overview of text-to-speech systems," in Advances in speech signal processing, eds. Furui, S. and Sondhi, M., New York: M. Dekker, bibtex: allen_overview_1992. 1

Megerdoomian, K. (2000), "Persian Computational Morphology: A unification-based approach," in NMSU, CLR, Memoranda in Computer and Cognitive Science Report. 2

Torkkola, K. (1993), "An Efficient Way To Learn English Grapheme-To-Phoneme Rules Automatically," pp. 199 – 202 vol.2. 1

اسلامی، محرم، شریفی آتشگاه، مسعود، احمدی‌نیا، زهرا، بهرامی راد، علی، و زندی، طاهره ،(۱۳۸۳a) "تبدیل رایانه‌ای متن به گفتار فارسی (گویا)،" در اولین کارگاه پژوهشی زبان فارسی و رایانه، دانشگاه تهران. ۲

اسلامی، محرم، شریفی آتشگاه، مسعود، علیزاده لمجیری، صدیقه، و زندی، طاهره ،(۱۳۸۳b) "واژگان زایای زبان فارسی،" در اولین کارگاه پژوهشی زبان فارسی و رایانه، دانشگاه تهران. ۲

اسلامی، محرم (۱۳۸۱)، "دشواری‌های پردازش رایانه‌ای خط فارسی،" نشر دانش، ۲۸ - ۳۲. ۳

اسلامی، محرم و علیزاده‌لمجیری، صدیقه (۱۳۸۸)، "ساختار تصریفی کلمه در زبان فارسی،" زبان و ادب فارسی، ۱ - ۱۸. ۲

ثمره، یدالله (۱۳۷۸)، آواشناسی زبان فارسی. مرکز نشر دانشگاهی، ویرایش دوم. ۳

شمس‌فرد، مهرنوش (۱۳۸۸)، :STeP۱" تهیه متن معیار برای زبان فارسی،" گزارش طرح تحقیقی، آزمایشگاه پردازش زبان طبیعی، دانشگاه شهید بهشتی. ۲

صیادیان، ابوالقاسم و نصیرزاده، مجید (۱۳۷۵)، "تجربه‌ای در مدل‌سازی زبان فارسی برای یک سیستم تبدیل متن به گفتار،" در دومین کنفرانس سالانه انجمن کامپیوتر ایران، صفحات ۱۰۵ - ۱۱۱. ۱