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

طراحی ابزار پارسر زبان فارسی

احمد استیری، محسن کاهانی، رضا سعیدی و احسان عسگریان
ahmad.estiri@stu.um.ac.ir, kahani@um.ac.ir, reza.saeedi@stu.um.ac.ir, asgarian@alum.sharif.edu
آزمایشگاه فناوری وب، دانشگاه فردوسی مشهد

چکیده

به موازات پیشرفت و تحولات نظری در زبان‌شناسی جدید، روش‌های تحلیل متون و دستورات زبان بوسیله‌ی رایانه نیز تحول یافته است. منظور از گرامر هر زبان، در دست داشتن یک سری دستورات زبانی قابل فهم برای رایانه است که به کمک آنها بتوان اجزای نحوی یک جمله را به طور صحیح تفکیک نمود. تجزیه و تحلیل جمله و شکستن آن به اجزای تشکیل‌دهنده مانند گروه‌های اسمی، فعلی، قیدی و غیره توسط ابزاری به نام پارسر صورت می‌گیرد که نقش اساسی در طراحی و یا افزایش دقت سایر ابزارهای پردازش متن دارد.

پارسر طراحی شده برای زبان فارسی از ساختار لغات، موقعیت و ترتیب لغات در جمله، حروف یا عبارات قبل و بعد از آنها و نوع لغات، درخت نحوی یا پارسینگ را برای جملات متن تشکیل می‌دهد. در واقع عملیات پارسینگ با توجه به ریخت‌شناسی (مطالعه ساختار و حالت-های مختلف یک کلمه) و همچنین دستورات نحوی گرامر زبان فارسی صورت می‌گیرد. بدیهی است هر چقدر نگارش بکار رفته در جملات و همچنین رعایت علائم سجاوندی طبق اصول و با دقت بیشتری صورت گرفته باشد، عملیات پارسینگ با کیفیت بهتری صورت خواهد گرفت و اجزای تشکیل‌دهنده‌ی جمله با عملیات کمتر و ساده‌تری برچسب زده خواهند شد.

کلمات کلیدی

پردازش زبان طبیعی، زبان فارسی، پارسر، درخت تجزیه، دستورات گرامری، ریخت شناسی

۱- مقدمه

پردازش متن از جمله مسائل اساسی در حوزه هوش مصنوعی و شناخت رایانشی است که در چند دهه اخیر، توجهات گسترده‌ای را در قالب‌های عدیده به خود معطوف کرده است.

پردازش متون زبان فارسی در سطوح چهارگانه‌ی آوایی، ساخت‌واژی، نحو و معنایی و همچنین در حوزه‌های کاربردی متعددی امکان‌پذیر می‌باشد. پردازش متون در سطح ساخت‌واژه و نحو منجر به طراحی ابزاری به نام پارسر می‌گردد. پارسرها با بهره‌گیری از دستورات گرامری زبان به تفکیک جملات متون به اجزای تشکیل‌دهنده‌ی آن، مشخص کردن نقش هر عبارت و لغت در متن و همچنین تشکیل درخت تجزیه برای جملات متن می‌پردازند.

پارسر نقش پایه‌ای و مهمی را در بهبود ابزارهای پردازش متن ایفا می‌کند. به عنوان مثال جهت تقویت الگوریتم‌های وابسته به برچسب‌زن نحوی لغات (POS tagger) و برچسب‌زن معنایی لغات (SRL) علاوه بر نقش‌های کلمات، وابستگی‌های کلمات به لحاظ نقشی در جمله نیز باید مشخص گردند. گرامر هر زبان، مجموعه قوانینی است که ویژگی‌ها و استعدادهای آن زبان را نشان می‌دهد. دستورات گرامری بکار گرفته شده در پارسر زبان فارسی در سیستم‌هایی نظیر سیستم‌های بازشناسی گفتار، سیستم‌های بررسی صحت ساختاری جملات، ترجمه ماشینی، سیستم‌های خلاصه‌ساز و تمامی ابزارهای پردازش متن قابل استفاده خواهد بود.

در پردازش متون زبان طبیعی با زبان نوشتاری سر و کار داریم. این مسأله باعث می‌شود گرچه به جهت از دست دادن اطلاعات گویشی مانند لحن گوینده، آهنگ صدا، تاکید و مکث، با مشکلات و ابهاماتی مواجه شویم، ولی در مقابل با شکل محدودتر و با قالب دستوری مشخص‌تری از زبان کار می-کنیم.

در تلاش برای ساخت یک سیستم پردازش و درک متون فارسی با مسائل و مشکلاتی مواجه می‌شویم که بعضی از آنها در بیشتر زبان‌ها بروز کرده و برخی خاص زبان فارسی می‌باشند. همچنین برخی از این پیچیدگی‌ها به طبیعت زبان و نارسایی‌های دستورات زبان‌شناسی مربوط و برخی دیگر برخاسته از مشکلات ایجاد سیستم‌های هوش مصنوعی است [۱].

با توجه به موارد ذکر شده و از آنجایی که زبان فارسی نوعی از زبان‌های غیرساختیافته است با مشکلات بسیار بیشتری نسبت به سایر زبان‌ها مواجه خواهیم شد. متون غیرساختیافته، متونی هستند که پیش‌فرض خاصی در مورد قالب آنها نداریم و آنها را به صورت مجموعه‌ای مرتب از جملات در نظر می‌گیریم.