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

تعیین محدوده جملات فارسی به کمک درخت تصمیم گیری

سیده طاهره میرعمادیان[۱]، خشایار یغمایی[۲] و سید مرتضی سیف الله پور[۳]
[۱] دانشجوی کارشناسی ارشد الکترونیک دانشگاه سمنان، tmiremadian@yahoo.com
[۲] عضو هیات علمی دانشگاه سمنان، khyaghmaie@semnan.ac.ir
[۳] دانشجوی کارشناسی ارشد الکترونیک دانشگاه سمنان، moriseif@gmail.com

چکیده

این مقاله به بررسی تعیین محدوده جملات فارسی به کمک درخت تصمیم گیری می پردازد.تعیین محدوده جمله از جمله مراحل پیش پردازش در اکثر الگوریتمهای متن کاوی و پردازش زبان طبیعی می‌باشد. داده‌های این مقاله در حدود ۲۳۰۰۰ کلمه می باشد که به همراه برچسب‌هایشان از پیکره بی جن خان تهیه شده است .تعداد برچسب‌های این مجموعه به دلیل پرهیز از پیچیدگی زیاد ، به ۱۸ برچسب تقلیل داده شده است و برچسب جدید محدوده جمله که شامل ابتدا،انتها ،وسط و ابتدا/انتها می‌باشد به کلمات تخصیص یافته است. نتایج حاصل از شبیه سازی برخت تصمیم گیری داده‌های فوق به کمک نرم افزار weka **، برابر ۹۶,۹۸ % می‌باشد.**

کلمات کلیدی

محدوده جملات فارسی، متن کاوی، درخت تصمیم گیری

۱- مقدمه

تعیین محدود جمله یک مسئله اساسی برای بسیاری از کاربردهای پردازش زبان طبیعی[۱] مانند تجزیه[۲]، استخراج اطلاعات[۳]،ترجمه ماشینی[۴]،خلاصه سازی[۵] و ... می‌باشد.اگرچه قطعه بندی جمله می‌تواند از روی علایم نشانه گذاری مانند نقطه ، ویرگول و یا حتی افعال در زبان فارسی صورت گیرد ،اما همچنان ابهام‌های زیادی در متون وجود دارد که عمل شناسایی را سخت می‌کند.

از کارهای انجام شده در این زمینه می‌توان به [1,2,3] اشاره کرد که در زبان انگلیسی کار شده است.هم چنین در[4,5] به بررسي محدوده جمله در زبان هندی پرداخته شده است و در [6] نیز زبان ژاپنی را مورد بررسی قرار داده اند.در این کارها مسئله تعیین حدود جمله تبدیل به مسئله رفع ابهام از علایم نشانه گذاری شده است که با آن به صورت یک مسئله رده بندی برخورد شده است.

در زبان فارسی متاسفانه کارهای زیادی صورت نگرفته است و در این زمینه تنها می‌توان به [7,8] اشاره نمود.در [7] تنها به رفع ابهام از علامت نشانه گذاری نقطه (".") پرداخته شده است که آیا انتهای جمله است و یا جزئی از کلمات اختصاری و مخفف می‌باشد.در [8] نیز به بررسي ساختاري جمله و تعیین حدود جمله به کمک چندین روش پرداخته است که شامل بکارگیری روش تشخیص فعل در جمله ، استفاده از مدل چندتایی[۶] بر تعیین حدود جمله و استخراج خصیصه از جملات و به کارگیری شبکه عصبی به عنوان طبقه بندی کننده می‌باشد.

این مقاله با هدف بهبود دقت شناسایی محدوده جملات با استفاده از استخراج تعدادی ویژگی از هر کلمه و همسایگان آن در متن و همچنین استفاده از مدل های دوتایی و به کار گیری درخت تصمیم گیری به عنوان طبقه کننده، به این امر پرداخته است که در بخش‌های بعدی به طور مفصل بحث خواهد شد.

۲- اقسام جمله از لحاظ دستوری

به لحاظ دستوری ، جمله یک یا مجموع چند واژه است که بر روی هم پیام کاملی را از گوینده به شنونده می‌رساند.در زبان فارسی جمله‌ها از چند حیث دسته بندی می‌شوند.

---

[۱] Natural Language Processing (NLP)
[۲] Parsing
[۳] Information Extraction
[۴] Machine Translation
[۵] Abstraction
[۶] N-garm