به نام خدا

وزارت علوم، تحقیقات، و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران

(ایرانداک)

خلاصه گزارش برای پژوهشگران

طراحی سامانهٔ برچسب‌دهی به اجزای کلام برای متون فارسی

1397

الهام علایی ابوذر

معرفی پژوهش

زبان فارسی دارای پیچیدگی‌هایی است که مشکلاتی بسیاری را در مسیر برچسب‌گذاری رایانه‌ای اجزای واژگانی کلام ایجاد می‌کند. یکی از این پیچیدگی‌ها مربوط به شکل یکسان برخی از تکواژها است که باعث ابهام در متون فارسی می‌شود. همچنین در فارسی هم‌نگاره‌های بسیاری به دلیل پیچیدگی‌های موجود در ساخت‌واژهٔ فارسی، به وجود می‌آیند. بررسی کلی هم‌نگاره‌ها در پیکره‌های متنی موجود فارسی نشان می‌دهد که تعداد هم‌نگاره‌ها در پیکره‌ها قابل توجه است و می‌توان گفت، بیشتر هم‌نگاره‌ها فراوانی بالایی در پیکره‌ها دارند. اکثر این هم‌نگاره‌ها، در اثر یکسان بودن نمود نوشتاری تکواژ یای نکره، یای اسم‌ساز (اسم مکان، اسمی که دال بر شغل یا محافظت و دارندگی است، اسم‌معنی یا اشیا، تصغیر و تحبیب، اسم مصدر یا حاصل مصدر)، شناسهٔ دوم شخص مفرد و یای صفت‌ساز (صفت فاعلی و مفعولی، صفتی که دال بر نسبت است) و یای متصل به گروه اسمی به وجود آمده‌اند (علایی : ۱۳۹۵). سوال مطرح در پژوهش حاضر این است که آیا با رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی مختوم به «-ی»، که فراوانی بالایی در پیکره‌های متنی فارسی دارند، کارایی یک سیستم برچسب‌زنی

خودکار، افزایش میابد و در نهایت می‌توان سامانه‌ای طراحی کرد که عمل برچسب‌دهی خودکار را با در نظر گرفتن رفع ابهام از برچسب هم‌نگاره‌های اسمی و صفتی مختوم به «-ی» در فارسی، با کارایی بهتری انجام دهد؟

روش پژوهش

از آنجائی‌که یکی از اهداف این طرح پژوهشی، محک زدن امکان بهبود عملکرد ابزار «هضم» ، به عنوان یکی از ابزارهای موجود برچسب‌دهی به اجزای کلام در فارسی، از طریق به کاربردن نرم‌افزار رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی فارسی مختوم به «ی» است، ابتدا ضرورت رفع ابهام از برچسب نحوی این هم‌نگاره‌ها توضیح داده شد و در ادامه الگوهای مستخرج از پیکره جهت رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی فارسی مختوم به «ی» به صورت فهرست نمایش داده شد (مستخرج از طرح پژوهشی نگارنده تحت عنوان «رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی فارسی(۱۳۹۵)». سپس نرم‌افزار رفع ابهام از این هم‌نگاره‌ها معرفی شد و به منظور دستیابی به هدف پژوهش، یعنی « بررسی امکان بهبود بخشیدن به یکی از سیستم‌های موجود برچسب‌دهی اجزای کلام، به نام «هضم» ، از طریق به کار بردن نرم‌افزار مجهز به رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی مختوم به «ی» در فارسی»، ارزیابی‌های متنوعی انجام شد که مبنای آنها دو معیار کلی «صحت» و «معیار اِف» است و عملکرد آن جهت بهبود بخشیدن به ابزار برچسب‌زن «هضم» مورد بررسی قرار گرفت.

یافته‌ها

به منظور ارزیابی دقت برچسب‌دهی با در نظر گرفتن الگوهای رفع ابهام از هم‌نگاره‌های اسمی و صفتی مختوم به «ی» ، متنی که پیش از این به ابزار «هضم» به عنوان ورودی وارد شده بود و برچسب‌گذاری شده بود، این بار با در نظر گرفتن الگوهای مستخرج از بررسی هم‌نگاره‌های اسمی و صفتی مختوم به «-ی» در بافت نحوی در پیکره، مجدداً برچسب‌ها با لحاظ الگوهای مذکور مورد بازبینی قرار گرفت. در این آزمایش دقت برچسب‌زن اجزای کلام «هضم» به صورت کلی (در سطح تمامی برچسب‌ها) با استفاده از معیار صحت و در دو حالت «با الگوها» و «بدون الگوها» انجام شده است. سپس به بررسی تأثیر هر الگو در بهبود دقت برچسب‌زن پرداخته شده است. در نهایت این نتیجه به دست آمد که اگر تنها الگوهایی که تأثیر مثبت در برچسب‌زنی داشته‌اند را به برچسب‌زن اضافه کنیم

صحت (Accuracy) کلی برچسب‌زن ۹۵,۶۹۱ درصد می‌شود که ۱,۳۴ درصد نسبت به حالتی که از تمام الگوهای حساس به بافت نحوی استفاده شود، بالاتر است. در نهایت با لحاظ کردن نتیجهٔ اعمال بخش رفع ابهام از برچسب هم‌نگاره‌های اسمی و صفتی مختوم به «-ی» روی ابزار برچسب‌زن «هضم» ، سامانه‌ای برای برچسب‌گذاری اجزای واژگانی کلام تهیه شد.

نتیجه گیری و پیشنهادها

سوال مطرح در پژوهش حاضر این بود که آیا با رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی مختوم به «-ی»، که فراوانی بالایی در پیکره‌های متنی فارسی دارند، کارایی یک سیستم برچسب‌زنی خودکار، افزایش میابد و در نهایت می‌توان سامانه‌ای طراحی کرد که عمل برچسب‌دهی خودکار را با در نظر گرفتن رفع ابهام از برچسب هم‌نگاره‌های اسمی و صفتی مختوم به «-ی» در فارسی، با کارایی بهتری انجام دهد؟ برای پاسخ به سوالات مطرح در این پژوهش، ابتدا به مرور پیشینهٔ پژوهش و ذکر پژوهش‌های انجام شده در حوزهٔ برچسب‌دهی به اجزای کلام پرداخته شد. در فصل سوم علاوه بر معرفی یکی از سیستم‌های برچسب‌دهی خودکار به اجزای کلام، به نام «هضم» ، روش انجام پژوهش ذکر شد و به ارزیابی نرم‌افزار تهیه شده جهت رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی مختوم به «-ی» در فارسی نیز پرداخته شد. نهایتاً این نتیجه به دست آمد که اگر تنها الگوهای حساس به بافت نحوی که تأثیر مثبت در برچسب‌زنی داشته‌اند را به برچسب‌زن «هضم» اضافه کنیم، صحت (Accuracy) کلی برچسب‌زن ۹۵,۶۹۱ درصد می‌شود که ۱,۳۴ درصد نسبت به حالتی که از تمام الگوهای حساس به بافت نحوی استفاده شود، بالاتر است. بنابراین در تهیهٔ سامانهٔ برچسب‌دهی به اجزای کلام، این الگوها لحاظ شدند و گزینه‌ای تحت عنوان «رفع ابهام» در سامانه در نظر گرفته شد. در نهایت فصل چهارم به معرفی سامانهٔ تهیه شده (مبتنی بر سیستم برچسب‌دهی موجود به نام «هضم») جهت برچسب‌دهی خودکار اجزای کلام، پرداخته شده است.