نشریه پژوهش‌های زبان‌شناسی تطبیقی
سال دهم- شماره ۱۹- بهار و تابستان ۱۳۹۹

بررسی امکان افزایش صحت یک ابزار برچسب‌دهی به اجزای کلام در فارسی[۱]

الهام علایی‌ابوذر[۲]

چکیده

در این پژوهش به بررسی تأثیر رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی مختوم به «ی» در فارسی روی عملکرد یک سیستم برچسب‌دهی خودکار به اجزای کلام، پرداخته شده است. سیستم مورد مطالعه در پژوهش حاضر، سیستم «هضم» است. در پژوهش حاضر، نرم‌فزاری جهت رفع ابهام از برچسب نحوی هم‌نگاره‌های مذکور، تهیه شد که خود مبتنی بر الگوهای حساس به بافت نحوی است؛ این الگوها حاصل بررسی هم‌نگاره‌های مذکور در پیکرهٔ بی‌جن‌خان است. بنابراین، هضم با پیکرهٔ بی‌جن‌خان آموزش دیده شد. ارزیابی کلی نرم‌افزار تهیه شده جهت رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی مختوم به «–ی» در فارسی، نشان می‌دهد اگر تنها الگوهای حساس به بافت نحوی که تأثیر مثبت در برچسب‌زنی داشته‌اند را به ابزار برچسب‌دهی اضافه کنیم، صحت کلی برچسب‌زن، در مقایسه با حالتی که از تمام الگوهای حساس به بافت نحوی استفاده شود، ۱٫۳۴ درصد بالاتر است.

کلیدواژه‌ها: هم‌نگاره‌های اسمی و صفتی مختوم به «–ی»، سیستم «هضم»، الگوهای حساس به بافت نحوی، صحت برچسب‌زنی

۱. این مقاله مستخرج از طرح پژوهشی تحت عنوان «طراحی سامانه برچسب‌دهی به اجزای کلام برای متون فارسی» است که در پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) اردیبهشت ۱۳۹۷ انجام شده است.

۲- استادیار پژوهشگاه علوم وفناوری اطلاعات ایران (ایرانداک) ✉ | Elham_alaee2000@yahoo.com

تاریخ دریافت: ۹۷/۱۰/۰۸ تاریخ پذیرش: ۹۸/۰۶/۲۶

شناسه دیجیتال (DOI): 10.22084/RJHLL.2019.16614.1834

۱- مقدمه

پردازش متن به‌عنوان یکی از حوزه‌های مرتبط با زبانشناسی رایانشی، سابقه‌ای دیرینه در زبان فارسی دارد و آزمایشگاه‌های مختلفی در دانشگاه‌ها و سازمان‌ها برای پردازش متن شکل گرفته است. انجام بسیاری از عملیات خودکار بر روی زبان مانند ترجمه، خلاصه‌سازی، تصحیح املا و غیره، مستلزم استفاده از مجموعه‌ای از ابزارها جهت پیش‌پردازش و آماده‌سازی متون است. تهیهٔ این ابزارها به دو صورت انجام می‌پذیرد: دسته اول روش‌های وابسته به زبان هستند که براساس برخی قواعد ساختاری زبان انجام می‌شوند. روش‌های دیگر مستقل از زبان هستند و بیشتر از روش‌های آماری، پیکره‌های زبانی و روش‌های یادگیری ماشینی استفاده می‌کنند؛ البته در برخی موارد ترکیبی از هر دو روش مورد استفاده قرار می‌گیرد. طراحی و پیاده‌سازی این ابزارها برای زبان‌های مختلف به طرق مختلف و مخصوص زبان مربوطه صورت می‌گیرد. مهم‌ترین ابزارهای پردازش زبان طبیعی[۱] عبارتند از: ابزار تشخیص جمله[۲] (این ابزار با توجه به کاراکترهای جداکنندهٔ جملات، توانایی تشخیص جملات را در متن ورودی دارد)، ابزار جداساز[۳] (ابزاری برای شکستن یک متن بر اساس واحدهای بامعنی مانند کلمه، پاراگراف و نمادهای معنادار مانند «فاصله[۴]» است)، ابزار شناسایی موجودیت‌های نام‌دار[۵] (ابزاری برای تشخیص اسم‌ها و نوع آن‌ها اعم از اسامی افراد، اماکن و مقادیر عددی است)، شبکهٔ واژگانی[۶] (مجموعه‌ای از لغات و ارتباط معنایی میان آن‌ها است)، ابزار ریشه‌یابی[۷] (این ابزار برای ریشه‌یابی لغات و تشخیص نوع کلمه ساخته شده از آن ریشه مانند اسم مکان، اسم زمان، صفت فاعلی، مفعولی، به‌کار می‌رود)، ابزار تشخیص میزان شباهت[۸] (ابزاری برای تشخیص میزان شباهت میان دو عبارت بر اساس پارامترهای مختلف مانند نوع اسم‌های مشابه به‌کار رفته، استفاده از word-net و... است)، ابزار تشخیص دهندهٔ گروه‌های نحوی[۹] (این ابزار برای تشخیص گروه‌های اسمی، فعلی و ... در یک جمله مورد استفاده قرار می‌گیرد)، ابزار برچسب‌دهندهٔ نقش نحوی[۱۰] (این ابزار نقش‌های نحوی کلمات در جملات مانند فاعل، مفعول مستقیم، مفعول غیرمستقیم و ... را مشخص می‌کند)، ابزار نشانه‌گذاری (در پیکره)[۱۱] (ابزاری است برای ایجاد یک نمونه از یک آنتولوژی در یک سند ورودی)، ابزار تعیین هم‌مرجع‌ها[۱۲] (ابزاری برای تعیین مرجع اسمی یک اسم یا یک ضمیر در جملات است) و

1. Natural Language Processing (NLP)
2. Sentence recognition
3. Tokenizer
4. space
5. Named entity recognition
6. Word net
7. Stemming
8. Similarity recognition
9. Chunker
10. syntactic role labeler
11. Annotator
12. Co-reference resolution

ابزار برچسب‌گذاری اجزای واژگانی کلام[1] (ابزاری برای مشخص کردن نوع کلمات از قبیل اسم، صفت، قید، فعل و ... است (www.bigdata.ir).

همان‌گونه که ذکر شد، در زبان‌شناسی پیکره‌ای، برچسب‌گذاری اجزای کلام، عمل انتساب برچسب به کلمات تشکیل‌دهندهٔ یک متن یا یک پیکره است. این برچسب‌گذاری براساس مقولهٔ آن کلمه در متن (مانند «اسم»، «فعل»، «قید»، «صفت»، و غیره) صورت می‌پذیرد. برای این منظور مجموعه‌برچسب‌ها[2] مانند موارد زیر انتخاب می‌شود و به هر واژه در متن یک برچسب اختصاص داده خواهد شد.

ADJ	صفت
ADR	نقش‌نمای ندا
ADV	قید
CONJ	نقش نمای همپایگی
IDEN	شاخص
N	اسم
PART	جزء دستوری
POSNUM	صفت شمارشی پسین
POSTP	حرف اضافه پسین
PR	ضمیر
PREM	پیش‌توصیف‌گر
PRENUM	صفت شمارشی پیشین
PREP	حرف اضافه پیشین
PSUS	شبه جمله
PUNC	علامت نگارشی
SUB	نقش نمای وابستگی
V	فعل

به‌عنوان مثال، متن زیر، نمونه‌ای از متن برچسب‌خورده است که توسط یک ابزار برچسب‌دهی به اجزای کلام در فارسی، ایجاد شده است،:

متن ورودی:

مردم مازندران در اواسط مرداد ماه جشنی به نام نوروز ماه دارند وقتی که اولین محصول برنج زودرس رسید بعد از جمع‌آوری و درو با همان برنج غذا درست می‌کنند و درخارج از روستا جشن پایان کار می‌گیرند. این

1. Part Of Speech (POS) tagging
2. Tag set

مراسم دست مانند سیزده به در است و اعتقاد دارند که این روز را حتماً باید بیرون از روستا به سر برد در واقع این جشن یک نوع سپاس‌گزاری به درگاه خداوند است.

متن برچسب‌خورده:

مردم N/مازندران N/در P/اواسط Ne/مرداد N/ماه N/جشنی N/به P/نام N/نوروز N/ماه N/دارند V/ وقتی CONJ/که CONJ/اولین NUM/محصول N/برنج N/زودرس AJ/رسید V/بعد P/از P/جمع N/ آوری V/و CONJ/درو N/با P/همان DET/برنج N/غذا N/درست AJ/می‌کنند V/و CONJ/درP/خارج N/ از P/روستا N/جشن N/پایان Ne/کار N/می‌گیرندPUNC./ V/ این DET/مراسم N/دست N/ مانند ADVe/سیزده NUM/به P/در N/است V/و CONJ/اعتقاد N/دارند V/که CONJ/این DET/ روز N/را POSTP/حتما PUNC/" ADV/باید V/بیرون ADV/از P/روستا N/به P/سر N/برد V/در P/ واقع N/این DET/جشن N/یک NUM/نوع CL/سپاسگزاری N/به P/درگاه Ne/خداوند N/استV / . /PUNC

(البته ممکن است درصدی از خطا هم در برچسب‌دهی خودکار وجود داشته باشد)

برچسب‌گذاری اجزای واژگانی کلام، از پیش‌نیازهای بسیاری از فعالیت‌های حوزهٔ پردازش زبان طبیعی از جمله ترجمهٔ ماشینی، خطایابی، تبدیل متن به گفتار، بازیابی اطلاعات و کمک به مدل‌های آماری است (مگردومیان: ۲۰۰۴). همچنین در طراحی سیستم نمایه‌ساز ماشینی، یکی از بخش‌ها، طراحی زیرسیستم تحلیل واژگانی است. این زیرسیستم، متن را به واژه‌ها تفکیک می‌کند و ماهیت هر کلمه را تشخیص می‌دهد و تشخیص نوع واژه و شناسایی فعل‌ها، الفاظ و اصطلاح‌ها را در بر دارد. بنابراین سیستم برچسب‌دهی خودکار، ماهیت مقولهٔ کلمات را مشخص می‌کند تا بتوان در مراحل بعدی از این اطلاعات در جهت استخراج کلیدواژه‌ها یا هر نوع بازیابی اطلاعات از متن، استفاده کرد. سامانه‌های برچسب‌گذاری، به دلیل عدم اشراف کامل به قواعد ساخت‌واژی زبان، ممکن است در برخورد با کلمات دارای پیچیدگی‌های ساخت‌واژی، با مشکلاتی مواجه شوند (محسنی: ۱۳۸۷). زبان فارسی نیز 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. بعضی کلمات در پیکره‌های متنی ممکن است بیش از یک برچسب داشته باشند؛ زیرا کلمات در جایگاه‌های مختلف می‌توانند برچسب‌های واژگانی متفاوت داشته باشند. مانند موارد زیر:

الف. برچسب یای نکره : فردا آسمانی صاف در انتظار شهروندان عزیز خواهد بود.

ب. برچسب یای صفت‌ساز: وی مردی آسمانی بود.

همچنین در فارسی هم‌نگاره‌های[1] بسیاری به دلیل پیچیدگی‌های موجود در ساخت‌واژهٔ فارسی، به‌وجود می‌آیند (بی‌جن خان و مرادزاده، ۱۳۸۳؛ نقل از محسنی، ۱۳۸۷). هم‌نگاره‌ها کلماتی هستند که صورت نوشتاری

1. homographs

یکسان، اما منشاء، معنی یا تلفظ متفاوت دارند (فرهنگ لغت مریم وبستر[1]). در زبان‌هایی که ساخت‌واژهٔ پیچیده دارند، مانند زبان فارسی، هم‌نگاره‌های بسیاری ساخته می‌شوند. گویشوران فارسی به دلیل مجهز بودن به اطلاعات زبانی از قبیل اطلاعات مربوط به ساخت‌واژهٔ فارسی، ساخت واجی فارسی و ساخت نحوی فارسی، قادر به رفع ابهام از هم‌نگاره‌ها در بافت نحوی هستند، اما سامانه‌های پردازش متون فارسی، به دلیل عدم دسترسی کامل به چنین اطلاعات زبانی، با مشکلاتی در پردازش هم‌نگاره‌ها مواجه می‌شوند که یکی از این مشکلات اختصاص برچسب درست اجزای کلام به هم‌نگاره‌ها در متون فارسی است. بررسی کلی هم‌نگاره‌ها در پیکره‌های متنی موجود فارسی نشان می‌دهد که تعداد هم‌نگاره‌ها در پیکره‌ها قابل توجه است. 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 به‌وجود آمده‌اند (علایی، ۱۳۹۵). سؤال مطرح در پژوهش حاضر این است که رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی مختوم به «–ی»، که فراوانی بالایی در پیکره‌های متنی فارسی دارند، چه تأثیری روی کارایی یک سیستم برچسب‌زنی خودکار (مطالعهٔ موردی: سیستم برچسب‌دهی خودکار «هضم[2]»)، دارد؟

۲- پیشینهٔ پژوهش

در این بخش به ذکر نمونه‌هایی از تحقیقاتی که در زمینهٔ برچسب‌گذاری اجزای کلام و هم‌نگاره‌ها، انجام شده است پرداخته می‌شود.

ویلکس و استیونسون (۱۹۹۸) به توضیح سامانه‌ای می‌پردازند که عمل ابهام‌زدایی از همهٔ کلمات قاموسی در متن را انجام می‌دهد؛ در این راستا از منابع اطلاعات گوناگونی استفاده می‌شود که شامل: ترجیحات معنایی، تعریف‌ها و توصیفات مربوط به کلمات موجود در فرهنگ‌های لغت و برچسب‌های اجزای واژگانی کلام است. در این سامانه از الگوریتم یادگیری نیز استفاده شده است. دقت سامانهٔ پیشنهادی به بیش از ۹۲٪ می‌رسد که ابهام‌زدایی از تمام کلمات را انجام می‌دهد و محدود به نمونهٔ کوچکی نیست.

عاصی و حاجی عبدالحسینی (۲۰۰۰) سامانهٔ برچسب‌دهی نحوی را به‌عنوان طرحی پژوهشی در پژوهشگاه علوم انسانی و مطالعات فرهنگی، معرفی می‌کنند؛ سامانهٔ مذکور، خود، در خدمت تهیهٔ پیکره‌ای فارسی به نام«پایگاه دادگان زبان فارسی (FLDB)» قرار می‌گیرد و می‌توان گفت اولین تلاش برای تهیهٔ سامانه‌ای جهت برچسب‌دهی نحوی زبان فارسی در نظر گرفته می‌شود. عاصی و حاجی عبدالحسینی به بررسی روشی می‌پردازند که شوتز (۱۹۹۵) برای برچسب‌دهی نحوی در متون انگلیسی به‌کار برده است و هدف آن‌ها بررسی

1. Merriam Webster

۲. «هضم» ابزاری است جهت پردازش زبان فارسی با استفاده از زبان برنامه‌نویسی پایتون که برای پیش‌پردازش‌هایی چون نرمال‌سازی متن، تقطیع جملات و واژه ها، ریشه یابی واژه ها، تحلیل صرفی واژه ها و تجزیهٔ نحوی جمله مورد استفاده قرار می‌گیرد. http://www.sobhe.ir/hazm

اعمال همان روش در فارسی است. آن‌ها معتقدند که با بهره‌گیری از محاسبات دقیق‌تر، می‌توان همان روش شوتز (۱۹۹۵) را در فارسی استفاده کرد.

فلیپ و زامورانو (۲۰۰۰) یک سیستم رفع ابهام از برچسب اجزای کلام و تقطیع‌کنندهٔ بخشی به نام Latch که برای زبان اسپانیایی تهیه شده است، معرفی می‌کنند. در این سیستم chunk ها شناسایی شده و در فرآیند رفع ابهام، ارجاع به آن‌ها مانند کلمات معمولی است. همچنین در این سیستم، جملات ساده‌سازی می‌شوند تا سیستم رفع ابهام‌کننده بتواند به تفسیر یک chunk همانند یک تفسیر یک کلمه بپردازد. تعامل دو سیستم رفع ابهام از برچسب نحوی کلمات و سیستم تقطیع‌کنندهٔ بخشی، به طور قابل توجهی منجر به کاهش سعی و تلاش مورد نیاز برای نوشتن قواعد می‌شود؛ علاوه بر این، روش پیشنهادی منجر به بهبود کارایی و نتایج می‌شود.

ریبرو و همکاران (۲۰۰۲) به ارائهٔ پیشرفت یک سیستم رفع ابهام ساخت‌واژی- نحوی (یا به عبارتی، سیستم برچسب‌گذاری اجزای کلام) می‌پردازند که به‌عنوان جزئی از یک سیستم تبدیل متن به گفتار برای زبان پرتغالی استفاده می‌شود. در فرآیند بهبود بخشیدن به فرآیند برچسب‌گذاری، دو رویکرد با هم مقایسه شده است: ۱- رویکردی مبتنی بر احتمالات و ۲- رویکردی تلفیقی. غیر از مقایسهٔ دو رویکرد مذکور، همچنین، این پژوهش به بررسی تأثیرات طبقات مختلف خطاها در عملکرد سیستم کامل تبدیل متن به گفتار می‌پردازد.

ورونسو (۲۰۰۴) به مشکل چگونگی ارتقاء کیفیت سیستم برچسب‌دهی به اجزای کلام در متن می‌پردازد. تکنولوژی معرفی شده بر اساس مکانیزم‌های مبتنی بر تحقیق و تجربه است که هدف آن‌ها، بهبود برونداد سیستم آماری جهت برچسب‌دهی نحوی است. ابزار پیشنهاد شده تلفیقی است از سیستم‌های برچسب‌دهی آماری و مبتنی بر قاعده که در لایه‌های گوناگون نحوی و واژگانی جمله عمل می‌کند.

ایندربو و همکاران (۲۰۰۵) به معرفی یک سیستم خودکار رفع ابهام از معنای کلمات می‌پردازند که از برچسب‌دهی اجزای کلام و طبقهٔ کلمات به‌عنوان مشخصه‌های مجزا، بهره می‌گیرد. طبقات کلمات از طریق تخصیص‌دهندهٔ طبقه به کلمات، گرفته می‌شود که خود از طریق پردازش آماری زبان به‌دست می‌آید. نتیجهٔ به‌دست آمده نشان می‌دهد که مشخصه‌های حاضر در برچسب‌دهی اجزای کلام، کارایی سیستم خودکار رفع ابهام از معنای کلمات را کمی بهبود می‌بخشد. تلفیق طبقهٔ کلمات و برچسب‌دهی به اجزای کلام، صحت رفع ابهام را به‌طور قابل توجهی بالا نبرده است و به‌طور کلی، بالا بردن صحت رفع ابهام از معنای کلمات مستلزم مطالعات و تحقیقات بیشتر و بهره‌گیری از سیستم‌های رفع ابهام از مشخصه‌های موجود در طبقهٔ کلمات است.

مونتویو و همکاران (۲۰۰۵) دو روش رفع ابهام معنایی از کلمات بر اساس دو رویکرد اصلی: روش مبتنی بر دانش و روش مبتنی بر پیکره معرفی می‌کنند. فرضیهٔ آنها این است که رفع ابهام معنایی از کلمات مستلزم استفاده از چندین منبع دانش است تا از این رهگذر بتوان ابهام معنایی کلمات را برطرف کرد. منابع مذکور انواع گوناگونی دارند. مانند: اطلاعات آماری، اطلاعات زنجیره‌ای[1]، اطلاعات مبتنی بر الگو وغیره. رویکرد تحقیق،

1. Syntagmatic information

تلفیقی است از منابع گوناگون دانش که با استفاده از دو روش مذکور (روش مبتنی بر دانش و روش مبتنی بر پیکره) انجام می‌شود. تمرکز این پژوهش عمدتاً روی چگونگی تلفیق این روش‌ها و منابع اطلاعات برای دستیابی به نتایج مطلوب در رفع ابهام است.

ژائو و مارکوس (۲۰۰۹) مدلی جدید برای برچسب‌دهی بدون‌نظارت/ بی‌سرپرست معرفی می‌کنند که مبتنی بر تمایز زبانی میان عناصر طبقهٔ باز و بسته کلمات است. در این سیستم از اطلاعات کمتری نسبت به سیستم‌های قبلی استفاده می‌شود و همچنین از روش‌های محاسباتی بسیار ساده‌تر بهره می‌گیرد. با به‌کار بردن تکنیک‌های سادهٔ یادگیری زبان که مبتنی بر محاسبه است، واژگان طبقهٔ بسته (واژگان دستوری مانند حروف اضافه، ضمایر، حروف ربط و...) به‌عنوان درونداد وارد سیستم می‌شوند و سیستم حجم زیادی از واژگان با طبقهٔ باز (کلمات قاموسی) را یاد می‌گیرد و سپس، قواعد رفع ابهام از هر دو طبقه را یاد می‌گیرد. این سیستم در مقایسه با سیستم‌های قبلی برچسب‌دهی بدون‌نظارت/ بی‌سرپرست، تحت شرایط یکسان، ۲۰٪ کاهش خطا دارد.

کلیشینسکی و همکاران (۲۰۱۱) به توضیح مدلی پیچیده جهت رفع ابهام از برچسب اجزای کلام برای متون روسی می‌پردازند. روش معرفی شده بر اساس اطلاعات مربوط به باهم‌آیی نحوی کلمات روسی است. همچنین آن‌ها دربارهٔ روش ساختن چنین پیکره‌ای نیز توضیح می‌دهند.

پاکری و مایومدر (۲۰۱۶) به توصیف رویکردی به نام POSTWITA برای برچسب‌گذاری نحوی اجزای کلام برای متون مطبوعاتی اجتماعی ایتالیایی می‌پردازند. این مقاله به توصیف مدل پیشنهادی این تیم جهت برچسب‌گذاری نحوی اجزای کلام می‌پردازد. مدل پیشنهادی مدلی باسرپرست/ با نظارت مبتنی بر دانش است که این دانش را از پایگاه دادگان کسب کرده است.

محسنی و مینایی بیدگلی (۱۳۸۸) با در نظر گرفتن مسائل و مشکلاتی که در مسیر برچسب‌گذاری اجزای کلام در فارسی وجود دارد، ابتدا طرحی کلی برای برچسب‌گذاری خودکار با دقت بالا در زبان فارسی پیشنهاد می‌کنند. سپس تحلیل ساخت‌واژی و استفاده از آن را برای پوشش دادن تعداد زیادی از برچسب‌های پیکره با حفظ دقت بالا در برچسب‌گذاری کلمات مورد بررسی دقیق‌تر قرار داده و تأثیر وجود یک تحلیل‌گر ساخت‌واژی در سطح تصریف را بر برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی بررسی می‌کنند. آن‌ها معتقد هستند نتایج به‌دست آمده نشان از کارآیی بسیار مناسب این روش پیشنهادی در برچسب‌گذاری دارد.

قیومی (۱۳۹۵) از برچسب‌زنی آماری برای برچسب‌زنی مقوله دستوری در دو سطح دانه‌ریز و دانه‌درشت استفاده می‌کند. وی تأثیر کیفیت و میزان اطلاعات مقوله‌های دستوری بر تجزیهٔ خودکار را بررسی می‌کند. به همین منظور سه سناریو برای تجزیهٔ خودکار جملات پیشنهاد می‌کند. در سناریوی نخست، تجزیه‌گر ابتدا باید دادهٔ ورودی را برچسب‌دهی کرده و سپس جمله را تجزیه کند. در سناریوی دوم از یک برچسب‌زن خارج از تجزیه‌گر استفاده شده است و در سناریوی سوم، برچسب معیار واژه‌ها برای تجزیهٔ جملات مورد استفاده قرار گرفته است. نتایج علمی نشان می‌دهد که کیفیت برچسب مقولات دستوری بر کارایی تجزیه‌گر تأثیر دارد. همچنین با انجام آزمایش‌ها برای بررسی تأثیرگذاری میزان اطلاعات از نظر دانه‌درشتی یا دانه‌ریزی این نتیجه

به‌دست آمد که اطلاعات بیشتر درمورد ویژگی‌های صرفی- نحوی هر واژه بر کارایی تجزیه‌گر تأثیر مثبت دارد و این تأثیرگذاری نسبت به کیفیت برچسب مقولات دستوری خیلی بیشتر است.

علایی (۱۳۹۵) ابتدا فهرست مبسوطی از هم‌نگاره‌های اسمی و صفتی مختوم به «-ی» با تعریف تعداد ۱۰ پنجره، به عبارتی دیگر، ۱۰ کلمه قبل و بعد از هر هم‌نگارهٔ مختوم به «ی»، از پیکرهٔ بی‌جن‌خان (که پیکره‌ای است برچسب‌خورده) تهیه کرده است؛ از آنجائی‌که همهٔ کلمات موجود در چنین پیکره‌ای دارای برچسب نحوی می‌باشند، الگوهای حساس به بافت نحوی جهت رفع ابهام از برچسب نحوی هم‌نگاره‌های مختوم به «ی» استخراج شده است. سپس جهت بررسی صحت الگوهای مذکور، برنامهٔ ماشینی تهیه شده است که صحت الگوهای مستخرج از بررسی هم‌نگاره‌های مختوم به «ی» را با در نظر گرفتن تعداد هم‌نگاره‌های بررسی شده در پیکره که می‌توان قاعده را در مورد آن‌ها بررسی کرد، تعداد موارد منطبق با هر قاعده، درصد موارد منطبق با هر قاعده، تعداد موارد مغایر با هر قاعده (تعداد موارد نقض) و درصد موارد مغایر با هر قاعده می‌سنجد. در نهایت بررسی ماشینی صحت الگوها نشان می‌دهد، صحت بیش از نیمی از الگوها بالای ۷۰٪ است.

۳- روش پژوهش

از آنجائی‌که هدف پژوهش بررسی تأثیر به کاربردن نرم‌افزار رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی فارسی مختوم به «ی» روی عملکرد یک ابزار برچسب‌دهی به اجزای کلام است، ابتدا ضرورت رفع ابهام از برچسب نحوی این هم‌نگاره‌ها توضیح داده می‌شود: پسوند «ی» از نظر بررسی‌های پیکره‌ای حائز اهمیت است، زیرا شمار زیادی از هم‌نگاره‌ها در پیکره‌های متنی فارسی، هم‌نگاره‌هایی هستند که در اثر اضافه شدن این پسوند، چه تصریفی و چه اشتقاقی، به ستاک ساخته می‌شوند (علایی، ۱۳۹۵). اطلاق هم‌نگاره به چنین ساخت‌هایی تنها فارغ از بافت صورت می‌پذیرد، زیرا بافت نحوی باعث رفع ابهام از چنین هم‌نگاره‌هایی می‌شود. گویشوران باسواد فارسی در مواجهه با هم‌نگاره‌ها، به دلیل آشنایی با ساخت‌واژه، ساخت واجی و نحوی فارسی و هم‌آیندها [1]در فارسی، قادر به رفع ابهام از هم‌نگاره‌ها هستند و در واقع، توانایی کاربرد صحیح هم‌نگاره‌ها را در بافت نحوی دارند؛ حال آن‌که سامانه‌های پردازش متن اگر به این اطلاعات مجهز نباشند در پردازش متن، تخصیص برچسب صحیح به کلمات، ترجمه ماشینی و... دچار خطا می‌شوند. به منظور دستیابی به هدف پژوهش، یعنی «بررسی تأثیر به‌کار بردن نرم‌افزار مجهز به رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی مختوم به «ی» روی برچسب‌گذاری اجزای کلام»، ارزیابی‌های متنوعی انجام شده است که مبنای آنها دو معیار کلی «صحت[2]» و «معیار اِف[3]» است. در ارزیابی‌های انجام‌شده، «معیار صحت» در مواردی استفاده شده است که ارزیابی کلی نرم‌افزار در سطح تمامی برچسب‌ها مدنظر بوده و «معیار اف» نیز هنگامی به‌کار رفته است که هدف، ارزیابی جزئی دقت سامانه در سطح تک‌تک برچسب‌ها به‌صورت مجزا بوده است.

1. Collocation
2. Accuracy
3. F-Measure

برای پیاده‌سازی قوانین از زبان برنامه‌نویسی پایتون[1] استفاده شده است. با جستجو در برچسب‌های اجزای کلام اعمال‌شده به متن ورودی توسط برچسب‌زن هضم، عبارات دارای الگوهای مدنظر مشخص شد و وضعیت انطباق آن‌ها با قوانین ارائه‌شده مورد بررسی قرار گرفت. در این راستا، به دلایلی که متعاقباً توضیح داده خواهد شد، ابزار برچسب‌گذاری هضم با «پیکرهٔ بی‌جن‌خان» آموزش دیده شد. برنامه به این صورت عمل می‌کند که با پیمایش در کلیهٔ واژه‌های متن، هرگاه با واژه‌ای مواجه شود که به «ی» ختم می‌شود، تک‌تک الگوها را در مورد این واژه بررسی می‌کند. به این معنا که این واژه را به‌عنوان واژهٔ مختوم به «ی» در هر الگو در نظر می‌گیرد و بررسی می‌کند که آیا با الگوی ارائه‌شده همخوانی دارد یا خیر. به منظور آشنایی با نحوهٔ عملکرد برنامه، چگونگی پیاده‌سازی دو مورد از الگوها برای نمونه تشریح می‌شود:

۱- حرف اضافه (P) + (کمیت‌نما /QUA) +***اسم (N)***

مثال از پیکره:

- در (P) آبادانی (N) شیراز کوشیدند
- به (P) هیچ (QUA) ارتشی(N) نیز اجازه نمی‌دهد....

۲- فعل (V) + حرف عطف (CON) / ویرگول (،) (DELM) +***اسم (N)*** + حرف ربط «که»

مثال از پیکره:

..... انجام می‌داد (V) ، (DELM) دوره‌ای (N) که تهران اقتصاد روستایی داشت....

۳- اسم (N) + قید (ADV) +***صفت (ADJ)***

مثال از پیکره:

- ما در شرایط (N) کاملاً (ADV) آزادی (ADJ) زندگی می‌کنیم
- هوای (N) بسیار (ADV) سردی (ADJ) است

۴- کمیت‌نما «هیچ» / «هر» + (نوع، گونه) / (عدد (N- SING -CN)) + اسم (N) + حرف عطف (CON) +***اسم (N)***

مثال از پیکره:

- هر (QUA) نوع سلیقه‌ای (N) در جواهری مظفریان.....
- هیچ (QUA) گناه (N) یا (CON) اشتباهی (N) صورت نگرفته است.

۵- صفت‌های اشاره «این» / « آن» + (نوع، همه) + ***اسم (N)*** + حرف ربط «که»

مثال از پیکره:

1. Python

- آن نوع برنامه‌سازی (N) را نخواهیم داشت.
- این شانسی (N) که شما دارید....

یافتن نمونه‌های منطبق با این الگوها به این صورت است که مثلاً در الگوی ۱، آیا واژهٔ قبل از واژهٔ مختوم به «-ی»، حرف اضافه است؟ یا آیا واژهٔ قبل از واژهٔ مختوم به «-ی»، کمیت‌نمایی است که خود پس از یک حرف اضافه آمده است؟ اگر این‌گونه بود برچسب «اسم» به این واژه اختصاص داده می‌شود. به منظور ارزیابی دقیق تأثیر افزودن الگوهای استخراج‌شده به برچسب‌زن اجزای کلام، آزمایش‌های مختلفی بر اساس دو معیار «صحت» و معیار «اف» انجام شد؛ به ساده‌ترین شکل، معیار «صحت» را می‌توان این‌گونه تعریف کرد: از مجموع برچسب‌هایی که سیستم به واژه‌ها (و علائم نگارشی) اختصاص داده است، چند درصد صحیح هستند یا به عبارتی چند درصد با برچسب‌های درست، که توسط نیروی انسانی مشخص شده‌اند، انطباق دارند (ژورافسکی و مارتین: ۲۰۰۹). معیار «اف» نیز از ترکیب دو معیار دقت[1] که با فرمول زیر محاسبه می‌شود (ژورافسکی و مارتین: ۲۰۰۹، فرمول ۱۷-۱۴) :

$$\text{دقت (P)} = \frac{\text{تعداد برچسب‌های x که سیستم به درستی تشخیص داده است}}{\text{تعداد کل برچسب‌های x تشخیص داده شده توسط سیستم}}$$

و فراخوانی[2] که با فرمول زیر محاسبه می‌شود (ژورافسکی و مارتین: ۲۰۰۹) :

$$\text{فراخوانی (R)} = \frac{\text{تعداد برچسب‌های x که سیستم به درستی تشخیص داده است}}{\text{تعداد کل برچسب‌های x در داده‌ای که نیروی انسانی برچسب زده است}}$$

به‌دست می‌آید. بنابراین براساس دو معیار فوق، معیار «اف» به شکل زیر محاسبه می‌شود (ژورافسکی و مارتین: ۲۰۰۹، فرمول ۱۲-۴۲) :

$$\text{معیار اف}(F) = \frac{2x\ P\ x\ R}{P + R}$$

با توجه به اینکه برای استخراج الگوها از «پیکرهٔ بی‌جن‌خان» استفاده شده است، در این بخش از پژوهش نیز به دو دلیل مهم باید از همین داده به‌عنوان دادهٔ حاوی برچسب‌های نیروی انسانی جهت انجام ارزیابی‌ها و بررسی تأثیر افزودن الگوها استفاده کرد: دلیل اول این است که مجموعه‌برچسب‌های هضم با مجموعهٔ برچسب‌های پیکرهٔ بی‌جن‌خان متفاوت است، بنابراین نمی‌شود مستقیماً برچسب‌زنی با هضم را با برچسب‌های پیکرهٔ بی‌جن‌خان سنجید و الگوهای استخراج‌شده را بر اساس برچسب‌زنی هضم اعمال کرد. دلیل دوم: پیکرهٔ استفاده شده برای پژوهش حاضر، پیکرهٔ بی‌جن‌خان است؛ حال آنکه مدل هضم متفاوت است. این مسئله

1. Precision
2. Recall

مشکل دیگری را برای ارزیابی برچسب‌زنی به‌وجود می‌آورد و آن این‌که اصولاً نباید داده‌ای که قبلاً برای آموزش سیستم به‌کار رفته است برای ارزیابی آن نیز به‌کار گرفته شود. راه حل این است که مدل جدیدی برای برچسب‌زن اجزای کلام هضم با استفاده از دادۀ بی‌جن‌خان آموزش داده شود. در نتیجه مجموعه‌برچسب‌های هضم در این حالت همان مجموعۀ برچسب‌های پیکرۀ بی‌جن‌خان خواهند بود که الگوها نیز پیش از این بر اساس آن‌ها تعریف شده‌اند.

با توجه به توضیحات فوق، از ارزیابی مرحله‌ای ده‌تایی[1] استفاده شده است؛ به این معنی که کل پیکرۀ بی‌جن‌خان به ۱۰ بخش تقسیم شد که ۹ بخش آن برای آموزش سیستم و یک بخش برای ارزیابی استفاده شد. به این ترتیب دادۀ آموزش[2] و دادۀ آزمون[3] کاملاً از هم مجزا هستند و هنگام ارزیابی مشکل تکراری بودن داده‌ها وجود ندارد. بنابراین، با توجه به اینکه قواعد مورد نظر بر اساس پیکرۀ بی‌جن‌خان استخراج شده بود، برچسب‌زن هضم با همین پیکره آموزش داده شد. برای این منظور برچسب‌زن با دادۀ آموزش (۰,۹ بی‌جن‌خان) آموزش داده شده و برچسب‌زنی بر روی دادۀ آزمون (۰,۱ بی‌جن‌خان) انجام و برچسب‌های هضم با برچسب‌های انسانی دادۀ آزمون مقایسه شده‌اند. 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 به‌عنوان ورودی وارد شده بود و برچسب‌گذاری شد، این بار با در نظر گرفتن الگوها مجدداً برچسب‌ها مورد بازبینی قرار گرفت. در این آزمایش دقت برچسب‌زن اجزای کلام هضم به‌صورت کلی (در سطح تمامی برچسب‌ها) با استفاده از معیار صحت و در دو حالت «با الگوها» و «بدون الگوها» انجام شده است. در حالت «با الگوها» تمامی الگوهای استخراج‌شده در طرح پژوهشی «*رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی فارسی* (علایی: ۱۳۹۵)» در هنگام برچسب‌زنی به‌کار گرفته شدند.

٤- یافته‌های پژوهش

جدول ۱، نتیجۀ بررسی دقت برچسب‌زن اجزای کلام هضم (که با پیکرۀ بی‌جن‌خان آموزش دیده است)، به‌صورت کلی (در سطح تمامی برچسب‌ها) با استفاده از معیار صحت و در دو حالت «با الگوها» و «بدون الگوها» را نشان می‌دهد:

جدول ۱: درصد صحت (Accuracy) کلی برچسب‌زن

بدون الگوها	با الگوها
۹۵,۶۹۰	۹۴,۳۵۱

1. fold cross validation-10
2. Train data
3. Test data

همان‌طور که ملاحظه می‌شود به‌کارگیری همهٔ الگوهای استخراج‌شده موجب کاهش ۱,۳۳۹ درصدی صحت برچسب‌زنی شده است. علت اختلاف کم میان دو حالت «بدون الگوها» و «با الگوها» دامنهٔ محدود اثرگذاری الگوها است. به‌عبارت دیگر قاعده‌های استخراج‌شده تنها در مورد واژه‌های مختوم به «ی» در دادهٔ آزمون، آن هم نه تمام آن‌ها بلکه مواردی که در بافت‌های خاصی قرار داشته باشند اعمال می‌شوند. این موارد در مجموع ۵,۱۷ درصد از کل واژه‌های دادهٔ آزمون را تشکیل می‌دهند و در نتیجه حداکثر میزان تأثیر الگوها می‌تواند همین مقدار باشد. در مرحله بعد، این بار تأثیر هر الگو در بهبود دقت برچسب‌زن مورد بررسی قرار گرفت؛ هدف از این آزمایش بررسی تأثیر تک‌تک الگوها بر بهبود دقت برچسب‌زن بوده است. برای محاسبهٔ تأثیر هر قاعده، معیار صحت بر اساس تمام مواردی که الگو اعمال شده است، یک بار بدون الگوها و یک بار با الگوها محاسبه شد. در این مرحله مشخص شد که برخی از الگوها دقت برچسب‌دهی را کمی بالا می‌برند. نتیجهٔ این بخش در جدول ۲ آورده شده است:

جدول ۲: درصد میزان تأثیر تک‌تک الگوها

قاعده	بدون الگوها	با الگوها	میزان تأثیر
rule01a	۰.۹۷۶۵۹	۰.۸۷۹۹۲	−۰.۰۹۶۶۸
rule01b	۰.۹۷۰۷۶	۰.۸۳۰۴۱	−۰.۱۴۰۳۵
rule02a	۰.۸۷۵۰۰	۰.۶۳۳۳۳	−۰.۲۴۱۶۷
rule02b	۱.۰۰۰۰۰	۰.۸۳۳۳۳	−۰.۱۶۶۶۷
rule03	۰.۹۰۰۰۰	۰.۴۲۵۰۰	−۰.۴۷۵۰۰
rule05a	۰.۸۳۷۸۴	۰.۴۰۵۴۱	−۰.۴۳۲۴۳
rule06a	۱.۰۰۰۰۰	۱.۰۰۰۰۰	۰.۰۰۰۰۰
rule06b	۰.۷۵۰۰۰	۱.۰۰۰۰۰	+۰.۲۵۰۰۰
rule07	۰.۸۸۲۳۵	۰.۴۳۵۲۹	−۰.۴۴۷۰۶
rule08	۰.۹۳۹۳۹	۱.۰۰۰۰۰	+۰.۰۶۰۶۰
rule08	۰.۹۵۷۴۵	۰.۷۲۳۴۰	−۰.۲۳۴۰۴
rule09a	۰.۹۴۶۵۰	۰.۹۳۶۳۱	−۰.۰۱۰۱۹
rule09b	۰.۸۷۱۴۶	۰.۷۶۱۱۳	−۰.۱۱۰۳۲
rule10a	۰.۹۳۰۴۳	۰.۷۳۹۱۳	−۰.۱۹۱۳۰
rule10b	۰.۹۵۴۷۵	۰.۷۷۳۷۶	−۰.۱۸۱۰۰
rule11	۰.۹۶۳۳۵	۰.۹۳۱۹۴	−۰.۰۳۱۴۱
rule12a	۰.۹۴۷۳۷	۰.۱۰۵۲۶	−۰.۸۴۲۱۱
rule12b	۱.۰۰۰۰۰	۱.۰۰۰۰۰	۰.۰۰۰۰۰
rule13	۰.۹۱۶۶۷	۰.۸۳۳۳۳	−۰.۰۸۳۳۳

rule14	۰.۹۴۵۴۱	۰.۴۵۴۷۲	−۰.۴۹۰۶۹
rule15a	۰.۹۵۵۷۸	۰.۵۷۳۳۱	−۰.۳۸۲۴۷
rule15b	۰.۹۴۵۲۱	۰.۸۰۸۲۲	−۰.۱۳۶۹۹
rule18	۰.۸۶۹۰۰	۰.۴۴۵۴۱	−۰.۴۲۳۵۸
rule20	۰.۹۵۶۰۴	۰.۶۸۱۳۲	−۰.۲۷۴۷۳
rule21a	۰.۹۴۴۰۶	۰.۸۱۱۱۹	−۰.۱۳۲۸۷
rule22a	۰.۹۴۸۲۰	۰.۵۷۳۲۰	−۰.۳۷۵۰۰
rule23a	۰.۹۶۴۸۲	۰.۸۱۴۰۷	−۰.۱۵۰۷۵
rule24	۰.۹۲۰۰۰	۰.۸۲۶۶۷	−۰.۰۹۳۳۳
rule25	۱.۰۰۰۰۰	۱.۰۰۰۰۰	۰.۰۰۰۰۰
rule26	۰.۹۷۸۲۶	۰.۹۳۴۷۸	−۰.۰۴۳۴۸
rule27	۰.۹۵۸۳۳	۰.۹۵۸۳۳	۰.۰۰۰۰۰
rule28	۰.۹۲۴۸۸	۰.۷۷۹۳۴	−۰.۱۴۵۵۴
rule29a	۰.۹۴۲۸۶	۰.۵۶۱۹۰	−۰.۳۸۰۹۵
rule30	۰.۹۸۱۸۲	۰.۴۵۴۵۵	−۰.۵۲۷۲۷
rule31	۱.۰۰۰۰۰	۰.۸۲۰۰۰	−۰.۱۸۰۰۰
rule32	۰.۷۸۹۴۷	۰.۵۲۶۳۲	−۰.۲۶۳۱۶
rule33a	۰.۹۴۱۵۶	۰.۴۳۷۲۳	−۰.۵۰۴۳۳
rule34a	۰.۹۴۸۸۴	۰.۶۰۹۳۰	−۰.۳۳۹۵۳
rule35	۱.۰۰۰۰۰	۱.۰۰۰۰۰	۰.۰۰۰۰۰
rule36	1.00000	0.80769	-0.19231

نهایتاً این نتیجه به‌دست آمد که اگر تنها الگوهایی که تأثیر مثبت در برچسب‌زنی داشته‌اند را به برچسب‌زن اضافه کنیم صحت (Accuracy) کلی برچسب‌زن ۹۵,۹۶۱ درصد می‌شود که ۱,۳۴ درصد نسبت به حالتی که از تمام الگوهای حساس به بافت نحوی استفاده شود، بالاتر است. در این پژوهش، تنها از کدهای هضم استفاده شده است و صحت خود ابزار هضم و مدل اصلی هضم چیز دیگری است.

۵- بحث دربارهٔ یافته‌های پژوهش

همان‌گونه که توضیح داده شد، هدف پژوهش حاضر بررسی میزان تأثیر رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی مختوم به «ی» در فارسی روی برچسب‌دهی خودکار اجزای کلام است. در این راستا نرم‌افزاری تهیه شد که مبتنی بر الگوهای حساس به بافت نحوی به دست آمده از پیکرهٔ بی‌جن‌خان است. از آنجائی‌که این الگوها مبتنی بر پیکرهٔ بی‌جن‌خان است، نمی‌توان آن‌ها را مستقیماً روی ابزار هضم پیاده‌سازی کرد؛ مدل استفاده شده در ابزار هضم متفاوت است و به دلایل ذکر شده در بخش ۳، ابزار هضم با پیکرهٔ

بی‌جن‌خان آموزش دیده شد (بنابراین، از مدل هضم استفاده نشده است). پیش‌فرض پژوهش این بود که به کار بردن چنین الگوهایی احتمال خطا در برچسب‌گذاری اجزای کلام را کاهش می‌دهد؛ مطالعهٔ انجام شده نشان می‌دهد که نمی‌توان از همهٔ ۳۶ الگوی پیکره–بنیاد تهیه شده استفاده کرد و استفاده از همهٔ الگوها موجب کاهش صحت برچسب‌گذاری می‌شود. بنابراین، تنها الگوهایی که تأثیر مثبت در برچسب‌گذاری داشته‌اند، لحاظ شده است و به ابزار هضم (آموزش دیده با پیکرهٔ بی‌جن‌خان) اضافه شده است. در این حالت، همان‌گونه که ذکر شد ۱٫۳۴ درصد بهبود عملکرد در برچسب‌گذاری، در مقایسه با حالتی که از تمام الگوها استفاده شود، مشاهده شد. به هر حال بهبود صورت پذیرفته است، حتی اگر به نظر کم باشد. در واقع هدف، بررسی تأثیر به‌کار بردن بخشی از دانش زبان‌شناسی در طراحی ابزار برچسب‌گذاری بوده است. می‌توان در آینده از دیگر حوزه‌های دانش زبان‌شناسی، مانند اطلاعات بافتی و هم‌آیندها، در برچسب‌گذاری خودکار اجزای کلام بهره برد و کارایی سیستم‌های برچسب‌گذاری اجزای کلام را افزایش داد.

٦- نتیجه‌گیری

سامانه‌های برچسب‌گذاری، به دلیل عدم اشراف به قواعد ساخت‌واژی زبان، در برخورد با کلمات دارای پیچیدگی‌های ساخت‌واژی، توان محدودی دارند. یکی از این پیچیدگی‌ها مربوط به شکل یکسان برخی از تکواژها است که باعث ابهام در متون فارسی می‌شود. در فارسی هم‌نگاره‌های بسیاری به دلیل پیچیدگی‌های موجود در ساخت‌واژهٔ فارسی، به‌وجود 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 به‌وجود آمده‌اند. سؤال مطرح در پژوهش حاضر این بود که رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی مختوم به «–ی»، که فراوانی بالایی در پیکره‌های متنی فارسی دارند، چه تأثیری روی عملکرد یک سیستم برچسب‌زنی خودکار، دارد؟ سیستم مورد مطالعه در پژوهش حاضر، سیستم «هضم» بود. در پژوهش حاضر، نرم‌افزاری جهت رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی مختوم به «–ی» در فارسی، تهیه شد که خود مبتنی بر الگوهای حساس به بافت نحوی است که بر اساس این الگوها می‌توان برچسب درست را به هم‌نگاره‌های مذکور اختصاص داد. ارزیابی کلی نرم‌افزار تهیه شده جهت رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی مختوم به «–ی» در فارسی، نشان می‌دهد اگر تنها الگوهای حساس به بافت نحوی که تأثیر مثبت در برچسب‌زنی داشته‌اند را به برچسب‌زن «هضم» آموزش‌دیده با پیکرهٔ بی‌جن‌خان، 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.

قدردانی

از جناب آقای مرتضی رضایی شریف‌آبادی که در برنامه‌نویسی پژوهش حاضر بنده را یاری رساندند، بسیار سپاس‌گزارم.

منابع

- علایی، الهام (۱۳۹۵). *بررسی ساخت‌واژی هم‌نگاره‌های اسمی و صفتی به منظور کمک به برچسب‌دهی «اسم» به کلیدواژه‌ها در پیکره‌های علمی*، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک).
- علایی، الهام (۱۳۹۵). *رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی فارسی*، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک).
- علایی، الهام و محمود بی‌جن‌خان (۱۳۹۲). «عمق خط فارسی. پژوهش‌های زبانی» (مجله سابق دانشکده ادبیات و علوم انسانی دانشگاه تهران)، دوره ۴، شماره ۱.
- قیومی، مسعود (۱۳۹۵). «بررسی مقایسه‌ای تأثیر برچسب‌زنی مقوله‌های دستوری بر تجزیه در پردازش خودکار زبان فارسی»، فصل‌نامهٔ پردازش علائم و داده‌ها، ۴ (۳۰). ۱۳۰-۱۲۱
- محسنی، مهدی (۱۳۸۷). *سیستم برچسب‌گذاری و ابهام‌زدایی خودکار اجزای کلام برای پیکره متنی زبان فارسی*، دانشگاه علم و صنعت. دانشکده مهندسی کامپیوتر
- محسنی، مهدی؛ مینایی بیدگلی، بهروز (۱۳۸۸). سیستم برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی. دو فصل‌نامه *پردازش علائم و داده‌ها*، ۲ (۱۲).
- Assi, M. and Haji Abdolhosseini, M. (2000). "Grammatical tagging of a Persian corpus", International journal of corpus linguistics. 5 (1): 69-82.
- Felipe, M.M., and Zamorano, J.P. (2000). POS disambiguation and Partial Parsing Bidirectional interaction. *LREC*
- https://github.com/sobhe/hazm
- http://www.merriam-webster.com
- Indrebo, K., Tao, J. and Trawicki, M. (2005). *Automatic word sense disambiguation (WSD) system*. Marquette University.
- Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2nd ed., Prentice Hall Series in Artificial Intelligence). Upper Saddle River, New Jersey: Pearson Education.
- Klyshinsky, E.S., Kochetkova, N.A., Litvinov, M.I., and Maximov, V.Yu. (2011). Method of POS disambiguation using information about words co-occurrence (for Russian). *Multilingual Resources and multilingual Applications*. 191.
- Megerdoomian, Karine (2004). *Developing a Persian part-of-speech tagger*. Proceedings of the 1st workshop on Persian language and computer, 99-105.

- Montoyo, A., Suarez, A., Rigau, G., and Palomar, M. (2005). "Combining knowledge – and corpus –based Word-Sense-Disambiguation methods". Journal of Artificial Intelligence Research. 23: 299-330.
- Pakray, P. and Majumder, G. (2016). NLP-NITMZ: part-of-speech tagging on Italian social media text using Hidden Markov Model. Shared Task On Postwita. 3RD Italian conference on computational linguistics. CLiC.
- Ribeiro, R., Oliveira, L., and Trancoso, I. (2002). Morphosyntactic disambiguation for TTS systems. *LREC*
- Vorontsov, A. (2004). Quality improvement of POS tagging in industrial text *processing systems.* National Conference on Modeling and Simulation, MS.
- Wilks, Y and Stevenson, M. (1998). Word sense disambiguation using optimized combinations of knowledge sources. *Proceedings of the 17th international conference on computational linguistics and the 36th annual meeting of the association for computational linguistics (COLING-ACL` 98).* Montreal, Canada. 2: 1398-1402.
- http://www.bigdata.ir
- http://www.sobhe.ir/hazm
- Zhao, Q. and Marcus, M. (2009). A simple unsupervised learner for POS disambiguation rules given only a minimal lexicon. *Proceedings of the conference on Empirical Methods in Natural Language Processing.* Association for computational linguistics, 2: 688-697.