

بهبود برچسب‌گذاری اجزای کلام با استفاده از نرم‌افزار رفع ابهام‌کننده از برچسب

هم‌نگاره‌های اسمی و صفتی مختوم به «-ی»

الهام علایی ابوذر *

چکیده

در پژوهش حاضر به این مسئله پرداخته شد که آیا با رفع ابهام از برچسب نحوی هم-نگاره‌های اسمی و صفتی مختوم به «-ی»، که فراوانی بالایی در پیکره‌های متنی فارسی دارند، کارایی یک سیستم برچسب‌زنی خودکار اجزای کلام، افزایش می‌یابد؟ سیستم برچسب‌زنی مورد مطالعه در پژوهش حاضر، ابزار «هضم» است. در پژوهش حاضر 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 هم‌نگاره مختوم به «-ی»، در پیکره بی‌جن‌خان (که پیکره‌ای است با برچسب اجزای کلام) تهیه شد؛ پس از بررسی هم‌نگاره‌های اسمی و صفتی مختوم به «-ی» در بافت نحوی، الگوهای حساس به بافت نحوی جهت رفع ابهام از برچسب نحوی هم‌نگاره‌های مذکور، استخراج شد؛ سپس، نرم‌افزاری جهت رفع ابهام از برچسب نحوی این هم‌نگاره‌ها، تهیه 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 برچسب‌زن «هضم» 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.

واژه‌های کلیدی: برچسب نحوی، هم‌نگاره‌های اسمی و صفتی مختوم به «-ی»، الگوهای

حساس به بافت نحوی، صحت برچسب‌زنی

۱- مقدمه

در زبان‌شناسی رایانشی، برچسب‌گذاری اجزای کلام، درواقع عمل انتساب برچسب به کلمات تشکیل‌دهنده یک متن یا یک پیکره است. این برچسب‌گذاری براساس مقوله آن کلمه در متن (مانند «اسم»، «فعل»، «قید»، «صفت»، و غیره) صورت می‌پذیرد. برچسب‌گذاری اجزای واژگانی کلام، از پیش‌نیازهای بسیاری از فعالیت‌های حوزه پردازش زبان طبیعی^۱ از جمله ترجمه ماشینی، خطایابی، تبدیل متن به گفتار، بازیابی اطلاعات و کمک به مدل‌های آماری است ([1] 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 مقوله 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 ([2] 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 : کشاورزی را دیدم.

ب. برچسب یای اسم‌ساز: وی چند سال کشاورزی را دنبال می‌کرد.

پ. برچسب یای شناسه دوم شخص مفرد: تو خود کشاورزی.

همچنین در فارسی هم‌نگاره‌های^۲ بسیاری به دلیل پیچیدگی‌های موجود در ساخت‌واژه

1 Natural language processing

2 homographs

فارسی، به وجود می‌آیند. هم‌نگاره‌ها کلماتی هستند که صورت نوشتاری یکسان، اما منشاء، معنی یا تلفظ متفاوت دارند ([3] فرهنگ لغت مریم وبستر^۱). در زبان‌هایی که ساخت‌واژه پیچیده دارند، مانند زبان فارسی، هم‌نگاره‌های بسیاری ساخته می‌شوند. گویشوران فارسی به دلیل مجهز بودن به اطلاعات زبانی از قبیل اطلاعات مربوط به ساخت‌واژه 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 اجزای کلام درست 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 تکیواژ 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 ([4] 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 (مطالعهٔ موردی: سیستم برچسب‌دهی خودکار «هضم»)، افزایش می‌یابد؟ «هضم» ابزاری است جهت پردازش زبان فارسی با استفاده از زبان برنامه‌نویسی پایتون^۲ که برای پیش‌پردازش‌هایی چون نرمالایز کردن متن، تقطیع جمله‌ها و واژه‌ها، ریشه‌یابی واژه‌ها، تحلیل صرفی واژه‌ها و تجزیهٔ نحوی جمله مورد استفاده قرار می‌گیرد.

[sobhe/hazm](#) ([5]).

۲- روش انجام پژوهش

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

1 Merriam Webster

2 Python

مجهز به رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی مختوم به «ی» در فارسی، ارزیابی‌های متنوعی انجام شده است که مبنای آنها دو معیار کلی «صحت»^۱ و «معیار اف»^۲ است. برای پیاده‌سازی قوانین از زبان برنامه‌نویسی پایتون استفاده شده است. با جستجو در برچسب‌های اجزای کلام اعمال شده به متن ورودی توسط برچسب‌زن هضم، عبارات دارای الگوهای مدنظر مشخص شد و وضعیت انطباق آن‌ها با قوانین ارائه شده مورد بررسی قرار گرفت. برنامه به این صورت عمل می‌کند که با پیمایش در کلیه 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 به عنوان واژه مختوم به «ی» در هر الگو در نظر می‌گیرد و بررسی می‌کند که آیا با الگوی ارائه شده همخوانی دارد یا خیر. به منظور آشنایی با نحوه عملکرد برنامه، چگونگی پیاده‌سازی دو مورد از الگوها برای نمونه تشریح می‌شود:

۱- زبان / لهجه + صفت (ADJ)

مثال از پیکره متنی:

...آشنایی کامل به چند زبان بین‌المللی (ADJ) از امتیازهای لازم...

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

«که»

مثال از پیکره متنی:

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

برای یافتن نمونه‌های منطبق با این الگوها بررسی می‌شود که مثلاً در الگوی ۱، آیا واژه قبل از واژه مختوم به «ی»، واژه «زبان» یا «لهجه» است؟ اگر اینگونه بود برچسب «صفت» به این واژه اختصاص داده می‌شود. به منظور ارزیابی دقیق تأثیر افزودن الگوهای استخراج شده 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

1 Accuracy
2 F-Measure

[illegible]

$$= \frac{2 \times P \times R}{P+R}$$

د: «اځا» معيار، فوق، معيار او داسې پر اساس بنديځ. داسې د:

$$= \frac{\text{تعداد پر خسته های داده ی تشخیصی}}{\text{تعداد کل پر خسته های داده ی آزمایشی که نیروی انسانی در دسترس است}}$$

و فراجاوانبی که با فرمول زیر محاسبه می شود

(p) وقت = $\frac{\text{تعداد کل برآیندهای X تشخیص داده شده توسط سیستم}}{\text{تعداد برآیندهای X که سیستم به درستی تشخیص داده است}}$

[illegible]

الگوهای استخراج‌شده در طرح پژوهشی «رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی فارسی» [7] (علائی: ۱۳۹۵) «در هنگام برچسب‌زنی به کار گرفته شدند. جدول ۱، نتیجه این بررسی را نشان می‌دهد:

جدول ۱: درصد صحت (Accuracy) کلی برچسب‌زن هضم

بدون الگوها	با الگوها
۹۵,۶۹۰	۹۴,۳۵۱

همانطور که ملاحظه می‌شود به کارگیری همه الگوهای استخراج‌شده موجب کاهش ۱,۳۳۹ درصدی صحت برچسب‌زنی شده است. در مرحله بعد، این بار تأثیر هر قاعده/الگو در بهبود دقت برچسب‌زن مورد بررسی قرار گرفت؛ هدف از این آزمایش بررسی تأثیر تک‌تک الگوها بر بهبود دقت برچسب‌زن بوده است. برای محاسبه تأثیر هر قاعده، معیار صحت بر اساس تمام مواردی که قاعده 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. قسمتی از نتیجه این بخش در جدول ۲ آورده شده است:

جدول ۲: میزان تأثیر تک‌تک الگوها

قاعده	بدون الگوها	با الگوها	میزان تاثیر
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
rule08DELM_hazm	۰.۹۵۷۴۵	۰.۷۲۳۴۰	-۰.۲۳۴۰۴
rule09a	۰.۹۴۶۵۰	۰.۹۳۶۳۱	-۰.۰۱۰۱۹
rule09b	۰.۸۷۱۴۶	۰.۷۶۱۱۳	-۰.۱۱۰۳۲

لازم به ذکر است به علت کمبود فضا، همه ۳۶ الگو در جدول ۲ آورده نشده است و تنها چند الگو به عنوان نمونه آورده شده است.

نهایتاً این نتیجه به دست آمد که اگر تنها الگوهایی که تأثیر مثبت در برچسب‌زنی داشته‌اند را به برچسب‌زن اضافه کنیم صحت (Accuracy) کلی برچسب‌زن ۹۵,۹۶۱ درصد می‌شود که ۱,۳۴ درصد نسبت به حالتی که از تمام الگوهای حساس به بافت نحوی استفاده شود، بالاتر است.

۳- نتیجه‌گیری

در پژوهش حاضر، ابتدا هم‌نگاره‌های اسمی و صفتی مختوم به «ی» در فارسی، در بافت نحوی مورد مطالعه قرار گرفت و الگوهای حساس به بافت نحوی جهت تخصیص برچسب نحوی صحیح به این هم‌نگاره‌ها، استخراج شد. سپس نرم‌افزاری تهیه شد که کاربرد الگوهای مذکور را با استفاده از معیارهای «صحت» و «اف» را در سیستم برچسب‌گذاری خودکار «هضم» می‌سنجید، در این پژوهش 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 شد. نهایتاً این نتیجه به دست آمد که اگر تنها الگوهایی که تأثیر مثبت در برچسب‌زنی داشته‌اند را به برچسب‌زن اضافه کنیم صحت (Accuracy) کلی برچسب‌زن ۹۵,۹۶۱ درصد می‌شود که ۱,۳۴ درصد نسبت به حالتی که از تمام الگوهای حساس به بافت نحوی استفاده شود، بالاتر است.

منابع

[48] Megerdooomian, Karine 2004. Developing a Persian part-of-speech tagger. *Proceedings of the 1st workshop on Persian language and computer*. Pp. 99-105.

[۴۹] محسنی، مهدی (۱۳۸۷). سیستم برچسب‌گذاری و ابهام‌زدایی خودکار اجزای کلام برای

پیکره متنی زبان فارسی. دانشگاه علم و صنعت. دانشکده مهندسی کامپیوتر

[۵۰] <http://www.merriam-webster.com>

[۵۱] علایی، ا. 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).

[52] <https://github.com/sobhe/hazm>

[53] 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.

[۵۴] علایی، ا. ۱۳۹۵. رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی فارسی. پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک).