

به نام خداوند مهربان



وزارت علوم، تحقیقات و فناوری  
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

طرح پژوهشی

**طراحی سامانهٔ برچسب‌دهی به اجزاء کلام برای متون فارسی**

مجری

الهام علایی ابوذری

اردیبهشت ۱۳۹۷

### حقوق: پژوهشگاه علوم و فناوری اطلاعات ایران

طرح پژوهشی «طراحی سامانه برچسب‌دهی به اجزاء کلام برای متون فارسی» پیرو قرارداد شماره ۳۳/۱۶۱۴۸ مورخ ۱۳۹۵/۱۲/۱۶ میان پژوهشگاه علوم و فناوری اطلاعات ایران (کارفرما) و خانم دکتر الهام علایی ابوذرا اجرا شده است. گزارش حاضر مستند این طرح است. این گزارش و تمامی حقوق مادی آن بر اساس «قانون حمایت از حقوق مؤلفان و مصنفان و هنرمندان» مصوب سال ۱۳۸۴ و اصلاحیه‌های بعدی آن و همچنین آیین‌نامه‌های اجرایی این قانون، متعلق به پژوهشگاه علوم و فناوری اطلاعات ایران است و هرگونه استفاده از تمام یا پاره‌ای از آن شامل نقل‌قول، تکثیر، انتشار، کاربرد نتایج، تکمیل و مانند آن به صورت چاپی، الکترونیکی یا وسایل دیگر تنها با اجازه کتبی پژوهشگاه امکان‌پذیر است. نقل‌قول در حد هزار واژه در انتشارات علمی مانند کتاب و مقاله با درج اطلاعات کامل کتابشناختی، نیازی به مجوز پژوهشگاه ندارد.



پایگاه

سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

پرونده ملی و فناوری اطلاعات ایران

شماره

شماره: ۳۴/۱۶۴۸  
تاریخ: ۱۳۹۷/۰۷/۱۵  
موضوع: قرارداد

شماره: ۳۴/۱۶۴۸  
تاریخ: ۱۳۹۷/۰۷/۱۵  
موضوع: قرارداد

### گواهی

انجام طرح پژوهشی:

عنوان: طراحی سامانه پرسبدهی به اجزاء کلام برای متون فارسی

مجری: دکتر الهام علایی ابوذر

همکار(ان):

شماره قرارداد: ۳۴/۱۶۴۸

تاریخ قرارداد: ۹۵/۱۲/۱۶

تاریخ شروع: ۹۵/۱۲/۱۶

تاریخ پایان: ۹۷/۲/۱

ناظران: دکتر مصطفی غاصبی، دکتر فرهاد شیرانی، میترا شمسی (ناظر فنی)

در پژوهشگاه علوم و فناوری اطلاعات ایران گواهی می‌شود. گزارش پایانی این طرح، بر پایه داوری در نشست ۲۸۵ (تاریخ ۹۷/۹/۱۱) شورای پژوهشی پژوهشگاه به تصویب رسیده است. لازم می‌دانم مراتب قدردانی خود را به مجری و ناظران این طرح تقدیم دارم.

دکتر زوی پرویزی  
رئیس شورای پژوهشی  
سرپرست معاونت پژوهش و آموزش

irandoc.ac.ir  
info@iranlib.ac.ir

پست: تهران، خیابان ولیعصر، پلاک ۱۳۳، طبقه ۳، کد پستی: ۱۹۵۱۴  
تلفن: ۰۲۱-۸۸۸۸۸۸۸۸  
فکس: ۰۲۱-۸۸۸۸۸۸۸۸

## سپاس‌گزاری

از اساتید گرامی، جناب آقای دکتر مصطفی عاصی و جناب آقای دکتر فرهاد شیرانی که نظارت این طرح پژوهشی را بر عهده داشتند و بسیار از نظرات عالمانه دو بزرگوار در جهت بهتر پیش بردن پژوهش بهرمند شدم، کمال تشکر را می‌نمایم. همچنین از آقای مرتضی رضایی شریف‌آبادی که در تهیه سامانه برچسب‌دهی به اجزای کلام، بنده را یاری رساندند، بسیار سپاس گزارم.

## طراحی سامانه برچسب‌دهی به اجزای کلام برای متون فارسی

مدیر طرح پژوهشی: الهام علایی ابوذر، عضو هیات علمی پژوهشکده علوم اطلاعات پژوهشگاه علوم و فناوری

اطلاعات ایران (ایرانداک)

نشانی: تهران، خیابان انقلاب، تقاطع فلسطین، شماره ۱۰۹۰

تلفن: ۶۶۹۵۱۴۳۰ (۰۲۱) دورنگار: ۶۶۴۶۲۲۵۴ (۰۲۱)

رایانامه: alayi@irandoc.ac.ir

## چکیده

در پژوهش حاضر به این مسئله پرداخته شد 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؟ 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» 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. این مسئله در تهیه سامانه برچسب‌گذاری اجزای کلام لحاظ شد و گزینه‌ای تحت عنوان «رفع ابهام» در سامانه در نظر گرفته شد. سامانه برچسب‌گذاری اجزای کلام، امکان وارد کردن متون گوناگون فارسی، برچسب‌گذاری مقوله‌ای کلمات تشکیل‌دهنده متون، مشاهده فهرست کلمات برچسب‌خورده همراه با فراوانی آن کلمات در متن، مشاهده فراوانی برچسب‌ها در متن، مشاهده فهرست اسم‌ها به ترتیب فراوانی آن‌ها در متن، رفع ابهام از برچسب برخی از هم‌نگاره‌های اسمی و صفتی فارسی و دریافت خروجی هر کدام از فهرست‌ها را فراهم می‌آورد.

## کلیدواژه‌ها:

هم‌نگاره‌های اسمی و صفتی مختوم به «-ی»، سیستم «هضم»، الگوهای حساس به بافت نحوی، صحت برچسب‌زنی، سامانه برچسب‌دهی به اجزای کلام در فارسی

## فهرست مطالب

| عنوان                                                                | شماره صفحه |
|----------------------------------------------------------------------|------------|
| فصل اول: مقدمه.....                                                  | ۱          |
| ۱-۱- مقدمه.....                                                      | ۲          |
| ۱-۲- سوال تحقیق.....                                                 | ۵          |
| ۱-۳- فرضیه تحقیق.....                                                | ۶          |
| فصل دوم: پیشینه پژوهش.....                                           | ۷          |
| ۱-۲- مقدمه.....                                                      | ۸          |
| ۲-۲- پیشینه پژوهش.....                                               | ۸          |
| فصل سوم: روش پژوهش.....                                              | ۱۵         |
| ۱-۳- مقدمه.....                                                      | ۱۶         |
| ۲-۳- معرفی ابزار پردازش متن «هضم».....                               | ۲۰         |
| ۳-۳- روش پژوهش.....                                                  | ۲۲         |
| فصل چهارم: معرفی سامانه برچسب‌دهی به اجزای کلام برای متون فارسی..... | ۵۹         |
| فصل پنجم: نتیجه‌گیری.....                                            | ۶۷         |
| فهرست منابع لاتین.....                                               | ۷۰         |
| فهرست منابع فارسی.....                                               | ۷۲         |

## فهرست جدول‌ها

- جدول ۳-۱- مثال برای معیار «صحت»..... ۳۷
- جدول ۳-۲- مثال برای معیار «اف»..... ۳۹
- جدول ۳-۳- مجموعهٔ برچسب‌های پیکرهٔ بی‌جن خان..... ۴۲
- جدول ۳-۴- برچسب‌های مشکل‌ساز در الگو..... ۴۳
- جدول ۳-۵- درصد صحت (accuracy) کلی برچسب‌زن «هضم»..... ۵۵
- جدول ۳-۶- دقت تک‌تک برچسب‌های «هضم» براساس معیار «اف»..... ۵۶
- جدول ۳-۷- میزان تاثیر تک‌تک الگوها..... ۵۷

# فصل اول

## مقدمه

## ۱-۱- مقدمه

مفهوم برچسب‌گذاری اجزای کلام<sup>۱</sup>، مفهومی برگرفته از زبان‌شناسی است که به طبقه‌نحوی کلمات مربوط می‌شود؛ به این صورت که ویژگی‌های ترکیبی یک کلمه در بافت/ ساخت نحوی مشخص می‌شود. در زبان‌شناسی پیکره‌ای، 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. برای این منظور مجموعه‌ای از برچسب‌ها<sup>۲</sup> مانند موارد زیر انتخاب می‌شود و به هر واژه در متن یک برچسب اختصاص داده خواهد شد.

|        |                  |
|--------|------------------|
| ADJ    | صفت              |
| ADR    | نقش نمای 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     |

<sup>۱</sup> Part Of Speech (POS) tagging

<sup>۲</sup> Tag set

طراحی سامانه برچسب‌دهی به اجزای کلام برای متون فارسی □ ۳

|        |                  |
|--------|------------------|
| PRENUM | صفت شمارشی پیشین |
| PREP   | حرف اضافه پیشین  |
| PSUS   | شبه جمله         |
| PUNC   | علامت نگارشی     |
| SUB    | نقش نمای وابستگی |
| V      | فعل              |

به عنوان مثال، نمونه‌ای از متن برچسب‌خورده به صورت زیر می‌باشد:

متن:

مردم مازندران در اواسط مرداد ماه جشنی به نام نوروز ماه دارند وقتی که اولین محصول برنج زودرس رسید بعد از جمع آوری و درو با همان برنج غذا درست می‌کنند و درخارج از روستا جشن پایان کار می‌گیرند. این 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/درخارج /N/از /P/روستا /N/جشن /N/پایان /Ne/کار /N/می‌گیرند /PUNC/. /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/است /V/. /PUNC/

بعضی کلمات 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.

اگر یک کلمه بیش از یک برچسب داشته باشد، نیاز به ابهام‌زدایی دارد.

برچسب‌گذاری اجزای واژگانی کلام، عملی کاربردی در بسیاری از حوزه‌های پیشرفته‌تر پردازش زبان طبیعی<sup>۱</sup> از جمله ترجمه ماشینی، خطایاب، تبدیل متن به گفتار، بازیابی اطلاعات، موتورهای جستجو و کمک به مدل‌های آماری است (مگردومیان: ۲۰۰۴). همچنین در طراحی سیستم نمایه‌ساز ماشینی، یکی از بخش‌ها، طراحی زیرسیستم‌ها است که خود بخش‌های گوناگونی دارد، یکی از این بخش‌ها، 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.

تاکنون از مدل‌ها و روش‌های بسیاری برای برچسب‌گذاری در زبان‌های مختلف استفاده شده است. این روش‌ها و مدل‌ها را می‌توان به دو دسته کلی تقسیم‌بندی کرد: دسته اول، رویکردهای آماری است که از به کار بردن روش‌های احتمالاتی برای پیکره‌های برچسب‌خورده<sup>۲</sup> بهره می‌جویند و دسته دوم، رویکردهای غیرآماری و مبتنی بر دانش<sup>۳</sup> / مبتنی بر قانون<sup>۴</sup> است که بر مبنای یادگیری ماشینی و دانش بشری استوار هستند، البته رویکرد سومی هم وجود دارد که در واقع تلفیقی از دو رویکرد آماری و مبتنی بر قانون است که برچسب‌گذاری تلفیقی<sup>۵</sup> خوانده می‌شود (مگردومیان: ۲۰۰۴). برخی از این روش‌های

<sup>۱</sup> Natural language processing

<sup>۲</sup> Tagged corpora

<sup>۳</sup> Knowledge-based

<sup>۴</sup> Rule-based

<sup>۵</sup> Hybrid tagging

گزارش شده شامل: مدل مخفی مارکوف<sup>۱</sup> (کارنیاک و همکاران: ۱۹۹۳ و کوپیک: ۱۹۹۳، نقل از محسنی: ۱۳۸۷)، سامانه‌های ماکزیمم بیشینه آنتروپی<sup>۲</sup> (راتناپارکی: ۱۹۹۶، نقل از محسنی: ۱۳۸۷)، برچسب-گذاری مبتنی بر تبدیل<sup>۳</sup> (بریل: ۱۹۹۴، ۱۹۹۳ و ۱۹۹۵، نقل از محسنی: ۱۳۸۷) و سامانه‌های مبتنی بر حافظه<sup>۴</sup> (دالمانس و همکاران: ۱۹۹۶، نقل از محسنی: ۱۳۸۷) است.

## ۲-۱- سوال تحقیق

سامانه‌های برچسب‌گذاری، به دلیل عدم اشراف به قواعد ساخت‌واژی زبان، در برخورد با کلمات دارای پیچیدگی‌های ساخت‌واژی، توان محدودی دارند. 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. همچنین در فارسی هم‌نگاره‌های بسیاری به دلیل پیچیدگی‌های موجود در ساخت‌واژه 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 هم-نگاره‌ها، در اثر یکسان بودن نمود نوشتاری تکواژ یای نکره، یای اسم‌ساز (اسم مکان، اسمی که دال بر شغل یا محافظت و دارندگی است، اسم معنی 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 برچسب-زنی 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؟

<sup>۱</sup> Hidden Markov model

<sup>۲</sup> Maximum entropy systems

<sup>۳</sup> Transformation-based tagger

<sup>۴</sup> Memory-based systems

### ۱-۳- فرضیه تحقیق

با توجه به تعداد بالای هم‌نگاره‌های اسمی و صفتی مختوم به «ی» در متون گوناگون فارسی و احتمال اینکه سیستم‌های خودکار برچسب‌دهی به اجزای کلام مجهز به نرم‌افزارهای رفع ابهام از برچسب‌های اجزای کلام نمی‌باشند، چنانچه سیستم موجود برچسب‌دهی خودکار اجزای کلام (مطالعه موردی: سیستم برچسب‌دهی خودکار «هضم») به نرم‌افزاری مجهز باشد که عمل رفع ابهام از برچسب نحوی اجزای کلام را انجام می‌دهد، عمل برچسب‌دهی با دقت بالاتری انجام خواهد شد و کارایی سامانه‌ای که در نهایت در این طرح پژوهشی، به عنوان برون‌داد، ارائه خواهد شد را افزایش خواهد داد.

در تحقیق حاضر ابتدا به مرور پیشینه پژوهش و ذکر پژوهش‌های انجام شده در حوزه برچسب‌دهی به اجزای کلام پرداخته می‌شود. در فصل سوم علاوه بر معرفی یکی از سیستم‌های برچسب‌دهی خودکار به اجزای کلام، به نام «هضم»، روش انجام پژوهش ذکر می‌شود که در این فصل به ارزیابی نرم‌افزار تهیه شده جهت رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی مختوم به «ی» در فارسی نیز پرداخته می‌شود. فصل چهارم معرفی سامانه تهیه شده (مبتنی بر سیستم برچسب‌دهی موجود به نام «هضم») جهت برچسب‌دهی خودکار اجزای کلام، می‌پردازد و در نهایت فصل پنجم، نتیجه‌گیری و فهرست منابع خواهد بود.

## فصل دوم

### پیشینه پژوهش

## ۲-۱- مقدمه

از آنجائی که رفع ابهام از برچسب نحوی هم‌نگاره‌ها به رفع ابهام معنایی از کلمات که گامی مهم در پردازش متن در حوزه زبان‌شناسی رایانشی است، نیز کمک خواهد کرد، در این فصل، علاوه بر بررسی پژوهش‌هایی در زمینه برچسب‌دهی خودکار متن انجام شده است، به تحقیقاتی که در زمینه رفع ابهام معنایی از کلمات با تاکید بر هم‌نگاره‌ها انجام شده است نیز اشاره می‌شود.

## ۲-۲- پیشینه پژوهش

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

۲-۲-۱- بینوت (۱۹۸۷) معتقد است که این مسئله که استفاده از نحو به منظور پشتیبانی از پردازش معنایی ضروری است، بدیهی است؛ برای اینکه بدانیم چه مقدار از نحو در این حوزه استفاده شود و اینکه چطور و چه زمانی آن را به کار بریم، مسئله‌ای مهم است. بینوت در این مقاله به بررسی راهکارهای استفاده شده در یک تحلیل‌گر معنایی در زبان فرانسه، به نام SABA، تهیه شده در دانشگاه لیگ بلژیک می‌پردازد. همچنین، در این مقاله، به طور خاص به بررسی فوائد دو فرآیند نحوی در پردازش معنایی پرداخته می‌شود: ۱- تجزیه و تحلیل به قطعات تشکیل‌دهنده که می‌تواند با پردازش معنایی همراه شود و ۲- رفع ابهام از برچسب اجزاء کلام که می‌تواند به عنوان مرحله پیش از پردازش عمل کند.

۲-۲-۲- رث و زلنکو (۱۹۹۸) الگوریتم یادگیری برخط و به کار بردن آن برای مسئله برچسب‌گذاری اجزاء کلام را ارائه می‌دهند. معماری ارائه شده در این پژوهش به نام SNOW، شبکه‌ای است از جداکننده‌های خطی در فضای مشخصه‌ها که در آن از الگوریتم به روز شده winnow استفاده شده است که الگوریتمی است با کارایی بالا. در این مقاله، معماری توضیح داده می‌شود که از این الگوریتم برای پیش‌بینی چند-طبقه‌ای (انتخاب برچسب یک کلمه) استفاده می‌شود. همچنین مطالعه مبسوط تجربی روی کارایی الگوریتم ذکر شده تحت شرایط گوناگون، به طور خاص انجام شده است و نتایج نشان می‌دهد که عملکرد

طراحی سامانه برچسب‌دهی به اجزای کلام برای متون فارسی □ ۹

الگوریتم ذکر شده قابل مقایسه با الگوریتم‌های شناخته شده برای برچسب‌دهی نحوی اجزاء کلام است.

۳-۲-۲- ویلکس و استیونسون (۱۹۹۸) به توضیح سامانه‌ای می‌پردازند که عمل ابهام‌زدایی از همه 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 اجزا واژگانی کلام است. در این سامانه از الگوریتم یادگیری نیز استفاده شده است. دقت سامانه پیشنهادی به بیش از ۹۲٪ می‌رسد که ابهام‌زدایی از تمام کلمات را انجام می‌دهد و محدود به نمونه کوچکی نیست.

۴-۲-۲- عاصی و حاجی عبدالحسینی (۲۰۰۰) سامانه برچسب‌دهی نحوی را به عنوان طرحی پژوهشی در پژوهشگاه علوم انسانی و مطالعات فرهنگی، معرفی می‌کنند؛ سامانه مذکور، خود، در خدمت تهیه پیکره‌ای فارسی به نام «پایگاه دادگان زبان فارسی (FLDB)» قرار خواهد گرفت و می‌توان گفت اولین تلاش برای تهیه سامانه‌ای جهت برچسب‌دهی نحوی زبان فارسی در نظر گرفته می‌شود. عاصی و حاجی عبدالحسینی به بررسی روشی می‌پردازند که شوتر (۱۹۹۵) برای برچسب‌دهی نحوی در متون انگلیسی به کار برده است و هدف آن‌ها بررسی اعمال همان روش در فارسی است. آن‌ها معتقدند که با بهره‌گیری از محاسبات دقیق-تر، می‌توان همان روش شوتر (۱۹۹۵) را در فارسی استفاده کرد.

۵-۲-۲- مارکز و همکاران (۲۰۰۰) طی پژوهشی، روش یادگیری استنتاجی درخت تصمیم آماری را برای پردازش زبان طبیعی و نوعاً در جهت رفع ابهام ساخت‌واژی-نحوی (برچسب‌دهی اجزاء کلام) به کار برده‌اند. فرآیند یادگیری از نوع بانظارت/با سرپرست است و از مدلی زبانی جهت رفع ابهام از برچسب اجزاء کلام بهره می‌گیرد؛ این مدل شامل مجموعه‌ای از درخت‌های تصمیم آماری است که نشان‌دهنده توزیع برچسب‌ها و کلمات در بافت‌های مربوط است. درخت‌های تصمیم ذکر شده مستقیماً در یک سیستم برچسب‌گذار استفاده

شده‌اند که باعث سهولت بخشیدن و سرعت بخشیدن به عمل برچسب‌دهی شده است و در نهایت، با پیکره (WSJ) wall street journal تست و ارزیابی شده است و دقت آن قابل توجه است. در این مقاله یک سیستم برچسب‌دهی توضیح داده شده است که قادر به استفاده از هر نوع اطلاعاتی از قبیل n-grams، محدودیت‌های زبانشناسی مربوط به نظام نوشتاری و... است. مارکز و همکاران در پایان مقاله نشان می‌دهند که با سیستم پیشنهادی، دقت بالاتری به دست می‌آید. همچنین برخی نتایج حاصل از به کار بردن این سیستم در پیکره اسپانیایی را نیز گزارش می‌کنند.

۶-۲-۲-

فلیپ و زامورانو (۲۰۰۰) یک سیستم رفع ابهام از برچسب اجزاء کلام و تقطیع‌کننده 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 پردازد. این تعامل دو سیستم رفع ابهام از برچسب نحوی کلمات و سیستم تقطیع‌کننده بخشی، به طور قابل توجهی منجر به کاهش سعی و تلاش مورد نیاز برای نوشتن قواعد می‌شود؛ علاوه براین، روش پیشنهادی منجر به بهبود کارایی و نتایج می‌شود.

۷-۲-۲-

ریرو و همکاران (۲۰۰۲) به ارائه پیشرفت یک سیستم رفع ابهام ساخت‌واژی-نحوی (یا به عبارتی، سیستم برچسب‌گذاری اجزاء کلام) می‌پردازند که به عنوان 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 مقایسه دو رویکرد مذکور، همچنین، این مقاله به بررسی تأثیرات طبقات مختلف خطاها در عملکرد سیستم کامل تبدیل متن به گفتار می‌پردازد.

۸-۲-۲-

ورونسو (۲۰۰۴) به مشکل چگونگی ارتقاء کیفیت سیستم برچسب‌دهی به اجزاء کلام در متن می‌پردازد. تکنولوژی معرفی شده در مقاله 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 سیستم‌های برچسب‌دهی آماری و مبتنی بر قاعده که در لایه‌های گوناگون نحوی و واژگانی جمله عمل می‌کند.

۹-۲-۲- ایندربو و همکاران (۲۰۰۵) به معرفی یک سیستم خودکار رفع ابهام از معنای کلمات می‌پردازند که از برچسب‌دهی اجزاء کلام و طبقه کلمات به عنوان مشخصه‌های مجزا، بهره می‌گیرد. طبقات کلمات از طریق تخصیص دهنده طبقه به کلمات، گرفته می‌شود که خود از طریق پردازش آماری زبان به دست می‌آید. نتیجه به دست آمده نشان می‌دهد که مشخصه‌های حاضر در برچسب‌دهی اجزاء کلام، کارایی سیستم خودکار رفع ابهام از معنای کلمات را کمی بهبود می‌بخشد. تلفیق طبقه کلمات و برچسب‌دهی به اجزاء کلام، صحت رفع ابهام را به طور قابل توجهی بالا برده است و به طور کلی، بالا بردن صحت رفع ابهام از معنای کلمات مستلزم مطالعات و تحقیقات بیشتر و بهره‌گیری از سیستم‌های رفع ابهام از مشخصه‌های موجود در طبقه کلمات است.

۱۰-۲-۲- مونتیو 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. فرضیه 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 (syntagmatic)، اطلاعات مبتنی بر الگو و غیره. رویکرد تحقیق، 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 مقاله عمدتاً روی چگونگی تلفیق این روش‌ها و منابع اطلاعات برای دستیابی به نتایج مطلوب در رفع ابهام است.

۱۱-۲-۲- مکی و همایون‌پور (۲۰۰۸) مدلی را برای ابهام‌زدایی از هم‌نگاره‌های فارسی معرفی می‌کند که از تزاروس و پیکره استفاده می‌کند. در مدل پیشنهادی از اطلاعات باهم‌آیی کلمات و کلمات موجود در بافت استفاده می‌شود. ۹۱/۴۶٪ از نمونه‌ای متشکل از ۱۵ هم‌نگاره فارسی، به درستی ابهام‌زدایی شده است. این مدل با سه مدل باسپرست و پیکره- بنیاد Naïve Bayes، Exemplar-based و Decision list مقایسه شده است؛ برخلاف روش‌های

طراحی سامانه برچسب‌دهی به اجزای کلام برای متون فارسی □ ۱۲

باسرپرست، روش پیشنهادی به داده‌ای برای آموزش نیازی ندارد و در مورد ابهام‌زدایی از کلمات غیررایج به خوبی عمل می‌کند. علاوه بر این، این روش قادر است بعضی از ابهامات ساخت‌واژی را نیز برطرف کند.

۱۲-۲-۲- ژائو و مارکوس (۲۰۰۹) مدلی جدید برای برچسب‌دهی بدون نظارت/ بی‌سرپرست معرفی می‌کنند که مبتنی بر تمایز زبانی میان عناصر طبقه‌باز 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 به کار بردن تکنیک‌های ساده یادگیری زبان که مبتنی بر محاسبه است، واژگان طبقه‌بسته (واژگان دستوری مانند حروف اضافه، ضمائر، حروف ربط و...) به عنوان درونداد وارد سیستم می‌شوند و سیستم حجم زیادی از واژگان با طبقه‌باز (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 بدون نظارت/ بی‌سرپرست، تحت شرایط یکسان، ۲۰٪ کاهش خطا دارد.

۱۳-۲-۲- کلشینسکی و همکاران (۲۰۱۱) در این مقاله به توضیح مدلی پیچیده جهت رفع ابهام از برچسب اجزاء کلام برای متون روسی می‌پردازند. روش معرفی شده بر اساس اطلاعات مربوط به باهم‌آیی نحوی کلمات روسی است. همچنین مقاله درباره روش ساختن چنین پیکره‌ای بحث می‌کند.

۱۴-۲-۲- پاکری و مایومدر (۲۰۱۶) به توصیف رویکردی به نام POSTWITA برای برچسب‌گذاری نحوی اجزاء کلام برای متون مطبوعاتی اجتماعی ایتالیایی می‌پردازند. این تحقیق در واقع به صورت تیمی انجام شده است و نام تیم پژوهشی NLP-NITMZ است که در مسابقه‌ای در حوزه پردازش زبان و با تمرکز بر چالش برچسب‌دهی نحوی اجزاء کلام شرکت کرده‌اند. این مقاله به توصیف مدل پیشنهادی این تیم جهت برچسب‌گذاری نحوی اجزاء کلام می‌پردازد. مدل پیشنهادی مدلی باسرپرست/ بانظارت مبتنی بر دانش است که این دانش را از پایگاه دادگان کسب کرده است.

۱۵-۲-۲- محسنی و مینایی بیدگلی (۱۳۸۸) با در نظر گرفتن مسائل و مشکلاتی که در مسیر برچسب‌گذاری اجزا کلام در فارسی وجود دارد، ابتدا طرحی کلی برای برچسب‌گذاری خودکار با دقت بالا در زبان فارسی پیشنهاد می‌کنند. سپس تحلیل ساخت‌وازی 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 ساخت‌وازی در سطح تصریف را بر برچسب‌گذاری اجزا واژگانی کلام در زبان فارسی بررسی می‌کنند. آن‌ها معتقد هستند نتایج به دست آمده نشان از کارایی بسیار مناسب این روش پیشنهادی در برچسب‌گذاری دارد.

۱۶-۲-۲- صدقی (۱۳۹۲)، برای رفع ابهام از معانی هم‌نویسه‌ها در متون فارسی، از الگوریتم فهرست‌تصمیم که روشی نظارتی و مبتنی بر پیکره است، استفاده کرده است. وی برای آموزش طبقه‌بند، تنها از کلمات پیرامون هم‌نویسه که هم‌رخداد نامیده می‌شوند، در پیکره آموزشی استفاده می‌کند. در این تحقیق، پیکره کوچکی برچسب‌گذاری شده و تلاش بر این بوده که با استفاده از پیکره‌های بدون برچسب که به سادگی و در اندازه‌های بزرگ در دسترس هستند، کارایی الگوریتم فهرست‌تصمیم که تحت تأثیر کمبود نمونه‌های برچسب‌دار است، بهبود یابد. نخست با الهام از tri-training یک روش نیمه‌نظارتی معرفی شده است که از افعال، اسامی و حروف برای خودگردان‌سازی آموزش، یعنی برچسب‌گذاری یک پیکره بزرگ، استفاده می‌کند. روش نیمه‌نظارتی tri-training روشی تکرارشونده است که کارایی آن به تخمین دقت طبقه‌بند مرحله قبل وابسته است و از این رو در صورت پایین بودن دقت برچسب‌گذاری طبقه‌بند مرحله قبل، ممکن است کارایی مطلوب به دست نیاید. استفاده از سه دیدگاه گرامری به بالا رفتن دقت کمک می‌کند. ارزیابی انجام شده نشان‌دهنده بهبود نتایج فراخوان تقریباً بین ۲ تا ۴۵ درصد، نسبت به فهرست‌تصمیم پایه هنگام کوچک بودن نمونه‌های آموزشی است. هم‌رخدادها ویژگی مهمی در آموزش فرضیه‌های یادگیری روش‌های رفع ابهام هستند. تشخیص هم‌رخدادهای موثر و تخصیص طبقه به هریک از آن‌ها در هر یک از روش‌های رفع ابهام به نحو خاصی انجام می‌شود. کوچک بودن پیکره آموزشی در روش‌های مبتنی بر پیکره بر این تشخیص‌ها تأثیر نامطلوب می‌گذارد. بنابراین از روش‌های

استخراج هم‌رخدادی به وسیله پیکره‌های بزرگ بدون برچسب می‌توان بهره برد. روش به کار رفته در این پژوهش علاوه بر توانایی تشخیص هم‌رخدادها می‌تواند جایگاه‌های هم-رخدادی نسبت به هم‌نویسه که موضوع تأثیرگذاری بر کارایی است را نیز تعیین کند. ارزیابی‌ها نشان می‌دهد استفاده از این هم‌رخدادها می‌تواند دقت فهرست تصمیم را عمدتاً بین ۲ تا ۳ درصد بالا ببرد. در سومین روش پیشنهادی از ایده استخراج هم‌رخدادهای پیشنهاد شده در روش پیشنهادی دوم استفاده شده تا دقت خودگردان سازی الگوریتم-tri-training افزایش یابد. به این ترتیب این امکان فراهم می‌شود تا بدون استفاده از سه دیدگاه گرامری نیز- که به معنای نیازمندی به برچسب گذاری گرامری پیکره بدون برچسب است- بتوان خودگردان‌سازی را بهبود داد.

۱۷-۲-۲- علایی (۱۳۹۵) 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 هم-نگاره مختوم به «ی»، در پیکره بی‌جن‌خان (که پیکره‌ای است برچسب‌خورده) تهیه شد؛ از آنجائی که همه 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 شد. سپس جهت بررسی صحت الگوهای مذکور، برنامه ماشینی تهیه شد 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 می‌سنجید. در نهایت بررسی ماشینی صحت الگوها نشان می‌دهد، صحت شا. از نم. از الگوها بالا، ۷۰٪ است.

## فصل سوم

### روش پژوهش

### ۳-۱- مقدمه

پردازش متن به عنوان یکی از زیرحوزه‌های فعال هوش مصنوعی، 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 زبان‌ها مانند ترجمه، خلاصه‌سازی، تصحیح املا و غیره، مستلزم استفاده از مجموعه‌ای از ابزارها جهت پیش‌پردازش و آماده‌سازی متون است. تهیه این ابزارها به دو صورت انجام می‌شود: دسته اول روش‌های وابسته به زبان هستند که براساس برخی قوانین نحوی و ساختاری زبان انجام می‌شوند. روش‌های دیگر مستقل از زبان هستند و بیشتر براساس پیکره‌های زبانی و با استفاده روش‌های یادگیری ماشینی صورت می‌گیرد. البته در برخی موارد ترکیبی از هر دو روش مورد استفاده قرار می‌گیرد. از این جهت طراحی و پیاده‌سازی این ابزارها برای زبان‌های مختلف به طرق مختلف و مخصوص زبان مربوطه صورت می‌گیرد. مهمترین ابزارهای پردازش زبان طبیعی در متون عبارتند از:

- **ابزاری برای تشخیص جمله<sup>۱۲</sup>:** این ابزار باید با توجه به کاراکترهای جداکننده جمله در خط و زبان، توانایی تشخیص جملات را در متن ورودی داشته باشد. برای ایجاد این ابزار باید ابتدا تمامی کاراکترها، نمادها و احیاناً قواعد دستوری که باعث شکسته شدن جملات می‌شوند، شناسایی گردند. با توجه به پایه بودن جمله در بسیاری از پردازش‌های زبانی، خروجی دقیق این ابزار از درجه اهمیت بالایی برخوردار است. از نمونه‌های انگلیسی آن می‌توان به OpenNLP، Stanford NLP، NLTK و Freeling اشاره کرد.

- **ابزار جداساز<sup>۱۳</sup>:** ابزاری برای شکستن یک متن بر اساس واحدهای با معنی مانند کلمه، پاراگراف و نمادهای معنادار مانند space (فاصله) و tab و ... است؛ لازمه ایجاد این ابزار در فارسی، جمع‌آوری واحدهایی است که در زبان فارسی به عنوان واحدهای مستقل معنایی شناخته می‌شوند؛ سپس بر اساس انتخاب هر کدام از این واحدها، متن بر اساس آن شکسته

<sup>12</sup> Sentence recognition

<sup>13</sup> Tokenizer

خواهد شد. از نمونه‌های انگلیسی آن می‌توان به Flex، JLex، JFLex، ANTLR، Ragel و Quex اشاره کرد.

- **ابزاری جهت شناسایی ماهیت اسم‌ها<sup>۱۴</sup>:** ابزاری برای تشخیص اسم‌ها و نوع آن‌ها اعم از اسامی افراد، اماکن، مقادیر عددی و ... است. برای تشخیص این که یک کلمه اسم است، راه‌های مختلفی وجود دارد که از جمله آن‌ها مراجعه به لغت‌نامه، مراجعه به word-net، در نظر گرفتن ریشه کلمه، استفاده از قواعد نحوی؛ ساخت‌واژه کلمه و ... می‌باشد. در این ابزار پس از تشخیص اسم‌ها با استفاده یک لغت‌نامه حاوی اسم افراد، مکان‌ها، مقادیر عددی و ... نوع اسم تشخیص داده می‌شود. از جمله نمونه‌های انگلیسی این ابزار می‌توان به Stanford NER و Illinois NER اشاره کرد.

- **شبکه کلمات<sup>۱۵</sup>:** مجموعه‌ای از لغات و ارتباط میان آن‌ها به لحاظ معنایی است. این مجموعه به عنوان یک مرجع در بسیاری از پردازش‌های زبانی مورد استفاده قرار می‌گیرد. از نمونه‌های انگلیسی آن می‌توان به Princeton Wordnet و EuroWordnet اشاره کرد.

- **ابزاری جهت ریشه‌یابی<sup>۱۶</sup>:** ابزاری برای ریشه‌یابی لغات و تشخیص نوع کلمه ساخته شده از آن ریشه (اسم مکان، اسم زمان، صفت فاعلی، مفعولی و...) است. معمولاً ریشه‌یابی لغات بر اساس قواعد ساخت‌وازی و سپس حذف پسوندها صورت می‌پذیرد. معروف‌ترین الگوریتم ریشه‌یابی در انگلیسی porter می‌باشد.

---

<sup>14</sup> Named entity recognition

<sup>15</sup> Word net

<sup>16</sup> Stemming

- **ابزاری جهت تشخیص میزان شباهت<sup>۱۷</sup>:** ابزاری برای تشخیص میزان شباهت میان دو عبارت بر اساس پارامترهای مختلف مانند نوع اسم‌های مشابه به کار رفته، استفاده از word-net و... است. در این ابزار پس از تشخیص نوع کلمات به کار رفته در یک جمله، بر اساس جایگاه آن کلمات در جمله، کلماتی که در جایگاه‌های یکسان قرار دارند، مورد مقایسه قرار می‌گیرند. از نمونه‌های انگلیسی آن می‌توان به Illinois WNSim و Illinois NESim اشاره نمود.
- **ابزاری جهت تشخیص دادن گروه‌های نحوی<sup>۱۸</sup>:** ابزاری برای تشخیص گروه‌های اسمی، فعلی و .... در یک جمله است. از جمله نمونه‌های انگلیسی آن می‌توان به Illinois Chunker اشاره کرد.
- **ابزار برچسب‌دهنده نقش نحوی<sup>۱۹</sup>:** ابزاری برای تشخیص نقش نحوی کلمه در جمله است. این ابزار یکی از مهم‌ترین نقش‌ها را در پردازش‌های زبانی بر عهده دارد. دقت در این ابزار بسیار حائز اهمیت است. این ابزار باید نقش‌های نحوی کلمات در جمله‌ها مانند فاعل، مفعول مستقیم، مفعول غیر مستقیم و .... را تشخیص دهد. از جمله نمونه‌های انگلیسی آن می‌توان به OpenNLP، Illinois SRL، Swirl و LTHSRL اشاره کرد. این ابزارها از الگوریتم پارسینگ charniak استفاده می‌کنند.

---

<sup>17</sup> Similarity recognition

<sup>18</sup> Chunker

<sup>19</sup> syntactic role labeler

- **ابزار نشانه‌گذاری ( در پیکره )<sup>۲۰</sup>:** ابزاری برای ایجاد یک نمونه از یک آنتولوژی در یک سند ورودی است. از ابزارهای موجود در انگلیسی می‌توان به Stanford Illinois Curator و Annotator اشاره کرد.
- **ابزار تعیین هم‌مرجع‌ها<sup>۲۱</sup>:** ابزاری برای تعیین مرجع اسمی یک اسم یا یک ضمیر در جملات است. استفاده از ضمائر به جای اسامی در زبان انگلیسی بسیار رایج است، اما در زبان فارسی این امر چندان رایج نیست. از نمونه‌های انگلیسی این ابزار می‌توان به Illinois Coreference package اشاره کرد.
- **ابزار برچسب‌گذاری اجزای واژگانی کلام:** ابزاری برای مشخص کردن نوع کلمات از قبیل اسم، صفت، قید، فعل و ... است. به عمل انتساب برچسب‌های واژگانی (اسم، صفت، قید، فعل و ...) به کلمات و نشانه‌های تشکیل‌دهنده یک متن برچسب‌گذاری کلمات (POS Tagging) گفته می‌شود؛ به صورتی که این برچسب‌ها نشان‌دهنده مقوله کلمات و نشانه‌ها در جمله باشد. البته برای متون فارسی، ما نیاز به انجام نرمال‌سازی<sup>۲۲</sup> متن هم داریم تا متون ما قبل از پردازش، یک‌دست و بدون اشکال گردند. برای نرمال‌سازی یک متن ابتدا باید همه نویسه‌های (کاراکترهای) متن با جایگزینی با معادل استاندارد آن، یکسان‌سازی گردند. متون مختلف ممکن است بسیار به هم شبیه باشند اما به دلیل تفاوت‌های ساده ظاهری از نظر ماشین متفاوت باشند؛ به همین دلیل ابتدا باید این تفاوت‌های ساده ظاهری برطرف گردد. همچنین اصلاحات دیگری نیز به منظور پردازش دقیق‌تر متون در این مرحله صورت می‌شود. برای رسیدن به این هدف، قبل از مقایسه متون، پیش‌پردازش‌هایی روی آن‌ها انجام می‌شود. طبیعتاً هر چه این پیش‌پردازش‌ها قوی‌تر باشد، نتایج حاصل از مقایسه متون قابل اطمینان‌تر خواهد بود. در پردازش رسم‌الخط زبان فارسی، با توجه به قرابتی که با رسم‌الخط عربی دارد، همواره در پردازش تعدادی از نگاره‌ها

<sup>20</sup> Annotator

<sup>21</sup> Coreference resolution

<sup>22</sup> Normalization

مشکلاتی وجود دارد که در اولین گام باید مشکلات مربوط به این نگاره‌ها را برطرف ساخت. علاوه بر این، اصلاح و یکسان سازی نویسه‌ی نیم‌فاصله و فاصله در کاربردهای مختلف آن و همچنین حذف نویسه‌ی «ـ» که برای کشش نویسه‌های چسبان مورد استفاده قرار می‌گیرد و مواردی مشابه برای یکسان سازی متون، از اقدامات لازم قبل از شروع فازهای مختلف می‌باشد.

در این فصل ابتدا به معرفی یکی از ابزارهای موجود برای پردازش متون فارسی تحت عنوان «هضم» پرداخته می‌شود و سپس روش پژوهش توضیح داده می‌شود.

### ۳-۲- معرفی ابزار پردازش متن «هضم»

«هضم» ابزاری است جهت پردازش زبان فارسی در پایتون. ویژگی‌های این ابزار شامل موارد زیر است

([sobhe/hazm](#)) :

- تمیز و مرتب کردن متن
- تقطیع جمله‌ها و واژه‌ها
- ریشه‌یابی واژه‌ها
- تحلیل صرفی جمله
- تجزیه نحوی جمله
- واسط استفاده از داده‌های زبان فارسی
- سازگاری با بسته [NLTK](#)
- پشتیبانی از پایتون نسخه ۲ و ۳
- تست مداوم کدها

نمونه‌ای از برخی از استفاده‌ها از این ابزار به صورت زیر آورده شده است:

## - نرمال‌سازی:

متن ورودی: «اصلاح نویسه‌ها و استفاده از نیم‌فاصله پردازش را آسان می‌کند»

اعمال نرمال‌سازی:

خروجی: «اصلاح نویسه‌ها و استفاده از نیم‌فاصله پردازش را آسان می‌کند»

## - ریشه‌یابی:

ورودی: «کتاب‌ها»

اعمال ریشه‌یابی:

خروجی: «کتاب»

ورودی: «می‌روم»

اعمال ریشه‌یابی:

خروجی: «رو# رفت»

## - برچسب‌گذاری اجزای واژگانی کلام:

متن ورودی: «ما بسیار کتاب می‌خوانیم»

اعمال برچسب‌گذاری اجزای واژگانی کلام:

خروجی:

[('V', 'می‌خوانیم'), ('N', 'کتاب'), ('ADV', 'بسیار'), ('PRO', 'ما'), ('P', 'ما')]

## - تشخیص دادن گروه‌های نحوی:

متن ورودی: « کتاب خواندن را دوست داریم »

اعمال chunking:

خروجی:

'[VP دوست داریم] [POSTP را] [NP کتاب خواندن]'

## ۳-۳- روش پژوهش

از آنجائی که هدف این بخش از پژوهش محک زدن امکان بهبود عملکرد ابزار «هضم» در زمینه برچسب‌دهی به اجزای کلام 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 می‌شود 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 می‌شود و در نهایت نرم‌افزار رفع ابهام از هم‌نگاره‌ها معرفی و عملکرد آن جهت بهبود بخشیدن به ابزار برچسب‌زن «هضم» بررسی می‌شود.

## ۳-۳-۱- ضرورت بررسی هم‌نگاره‌های اسمی و صفتی مختوم به «ی» در فارسی

پسوند «ی» از نظر بررسی‌های پیکره‌ای حائز اهمیت است، زیرا شمار زیادی از هم‌نگاره‌ها در پیکره-های 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<sup>۲۳</sup> در فارسی، قادر به رفع ابهام از هم‌نگاره‌ها هستند و در واقع، توانایی کاربرد صحیح هم‌نگاره‌ها را در بافت

<sup>23</sup> Collocation

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 خطاهای غیرقابل چشم‌پوشی می‌شوند.

### ۳-۲-۳- الگوهای حساس به بافت نحوی

الگوهای حساس به بافت نحوی جهت رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی مختوم به «ی» در فارسی در فهرست زیر خلاصه می‌شود؛ این الگوها پس از استخراج هم‌نگاره‌های اسمی و صفتی مختوم به «ی» در بافت نحوی از پیکره بی‌جن‌خان تهیه شده است (علایی: ۱۳۹۵). لازم به ذکر است اسم‌ها و صفت‌ها در الگوهای زیر همه مختوم به «ی» هستند که البته در برخی موارد «ی» تظاهر نوشتاری کسره اضافه است.

الگوی ۱. الف. حرف اضافه (P) + (کمیت‌نما / QUA) + اسم (N)

الگوی ۱. ب. حرف اضافه (P) + (کمیت‌نما / QUA) + اسم (N) + حرف عطف (CON) + اسم (N)

مثال‌ها:

- با (P) آیین‌نامه‌ای (N) روبرو شد....
- او از (P) سردی (N) و (CON) خشکی (N) و صبر و شکیبایی.....

الگوی ۲. الف. حرف اضافه (P) + لحاظ / شکل / صورت / نظر + اسم (N)

الگوی ۲. ب. حرف اضافه (P) + لحاظ / شکل / صورت / نظر + صفت (ADJ) + حرف عطف (CON) +

اسم (N)

مثال‌ها:

- به (P) لحاظ ارزشی (N) بین قله‌های تکیه زیرویمی و.....

- ظاهراً از (P) نظر فکری (N) و (CON) احساسی (N) بر دیگران.....

الگوی ۳. به عنوان + ( صفت عالی (ADJ-SUP) ) + اسم (N)

مثال:

- اسلام به عنوان دینی (N) جامع.....

- به عنوان جوان‌ترین (ADJ-SUP) کارگردانی (N) که.....

الگوی ۴. حرف اضافه (P) + اسم (N) + حرف اضافه (P)

مثال:

- برای (P) تاثیرگذاری (N) بر (P) اعضای اوپک.....

الگوی ۵. الف. دوره / دوران / زمان / سال / سال‌های / عصر / سن / سنین + اسم (N)

الگوی ۵. ب. حرف اضافه (P) + دوره / دوران / زمان / سال / سال‌های / عصر / سن / سنین + اسم (N)

مثال‌ها:

- دوره کهنسالی (N) ، دوره بسیار .....

- در (P) دوران جوانی (N) ممکن است.....

الگوی ۶. الف. کمیت‌نما «هیچ» / «هر» + (نوع، گونه) / (عدد (N- SING -CN) ) + اسم (N)

طراحی سامانه برچسب‌دهی به اجزای کلام برای متون فارسی □ ۲۵

الگوی ۶. ب. کمیت‌نما «هیچ» / «هر» + (نوع، گونه) / (عدد (N- SING -CN) + اسم (N) + حرف عطف (CON) + **اسم (N)**

مثال‌ها:

- هر (QUA) نوع آب و هوایی (N).....
- هر (QUA) یک (N- SING -CN) وظیفه (N) و (CON) تخصصی (N) داشته باشند

الگوی ۷. فعل (V) + حرف عطف (CON) + **اسم (N)** + حرف اضافه (P)

مثال:

- حتی یک مورد نمی‌توان سراغ گرفت (V) که (CON) شهروندی (N) به (P) اجبار و اکراه.....

الگوی ۸. فعل (V) + حرف عطف (CON) / ویرگول (،) (DELM) + **اسم (N)** + حرف ربط «که»

مثال:

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

الگوی ۹. الف. **صفت (ADJ)** + حرف عطف (CON) + صفت (ADJ)

الگوی ۹. ب. صفت (ADJ) + حرف عطف (CON) + **صفت (ADJ)**

مثال‌ها:

- موضع طبیعی (ADJ) یا (CON) قراردادی (ADJ) خود خارج....
- بر حاشیه‌اش رنگ‌های زرین (ADJ) و (CON) آبی (ADJ) جلوه‌گر بود

الگوی ۱۰. الف. اسم (N) + **صفت (ADJ)** + حرف عطف (CON) + اسم (N) + صفت (ADJ)

الگوی ۱۰. ب. اسم (N) + صفت (ADJ) + حرف عطف (CON) + اسم (N) + **صفت (ADJ)**

مثال‌ها:

- همان تغییر (N) اجتماعی (ADJ) و تغییر (CON) (N) فرهنگی (ADJ) لازم.....
- معاملات (N) سیاسی (ADJ) یا فشار (CON) (N) نظامی (ADJ) و امنیتی

الگوی ۱۱. خانم / آقای / القاب (دکتر، مهندس، استاد، سردار و...) + **اسم (N)**

مثال:

- مرحوم استاد فلسفی (N)

الگوی ۱۲. الف. شیخ + اسم (N) + **صفت (ADJ)**

الگوی ۱۲. ب. شیخ + **اسم (N)**

مثال‌ها:

- شیخ علی (N) یزدی (ADJ)
- شیخ طوسی (N) از جابر جعفری روایت کرده.....

الگوی ۱۳. قید (ADV-EXM-GEN) مانند / مثل / شامل / از قبیل / از جمله / چون / چو + **اسم (N)**

طراحی سامانه برچسب‌دهی به اجزای کلام برای متون فارسی □ ۲۷

مثال:

- مانند (ADV-EXM-GEN) قهوه‌ای (N) که کافئین آن را گرفته‌اند

الگوی ۱۴. اسم (N) + صفت (ADJ) + **صفت (ADJ)**

مثال:

- شرایط (N) نامطلوب (ADJ) آب و هوایی (ADJ)

الگوی ۱۵. الف. نقطه (.) (DELM) / نقطه ویرگول (؛) (DELM) / دونقطه (:) (DELM) + **اسم (N)**

الگوی ۱۵. ب. نقطه (.) (DELM) / نقطه ویرگول (؛) (DELM) / دونقطه (:) (DELM) + اسم (N) +

حرف عطف (CON) + **اسم (N)**

مثال‌ها:

- . (DELM) بادی (N) که یخ‌ها را آب می‌کند....

- . (DELM) فرد (N) یا (CON) گروهی (N) که آن‌ها را...

الگوی ۱۶. نقطه (.) (DELM) + **اسم (N)** + «را»

مثال:

- . (DELM) اسمی (N) را تصور کنید که....

الگوی ۱۷. ویرگول (،) (DELM) + **اسم (N)** + حرف اضافه (P)

طراحی سامانه برچسب‌دهی به اجزای کلام برای متون فارسی □ ۲۸

مثال:

- ، (DELM) ابزاری (N) برای (P) صلح و گفت‌وگوی....

الگوی ۱۸. عدد (N- SING -CN) + (مورد / نفر / واحد) + اسم (N)

مثال:

- یک (N-SING-CN) نوجوانی (N) است که او را به آموزشگاه....

الگوی ۱۹. عدد (N-SING-CN) + خانم / آقا / دختر / پسر / لقب + صفت (ADJ) + کسره اضافه (با نمود نوشتاری «ی» )

مثال:

- یک (N-SING-CN) دختر دانشجوی (ADJ) ۲۱ ساله....

الگوی ۲۰. اسم (N) + قید (ADV) + صفت (ADJ)

مثال:

- ما در شرایط (N) کاملاً (ADV) آزادی (ADJ) زندگی می‌کنیم

الگوی ۲۱. الف. صفت (ADJ) + (، ) (DELM) + صفت (ADJ)

قاعده ۲۱. ب. صفت (ADJ) + (، ) (DELM) + صفت (ADJ)

مثال‌ها:

- بخش‌های خارجی (ADJ) ، دریایی (DELM) ، نظامی (ADJ) ،
- ماهنامه‌ی بهداشتی (ADJ) ، دارویی (DELM) (ADJ) و اجتماعی

الگوی ۲۲. الف. اسم (N) + ویرگول ( ، ) (DELM) + اسم (N)

الگوی ۲۲. ب. اسم (N) + ویرگول ( ، ) (DELM) + اسم (N)

مثال‌ها:

- حق‌پرستی (N) ، پایداری (DELM) (N) ، مقاومت...
- شعار استقلال (N) ، آزادی (DELM) (N) ، حکومت اسلامی نبود

الگوی ۲۳. الف. صفت‌های اشاره «این» / «آن» + نوع، همه + اسم (N)

الگوی ۲۳. ب. صفت‌های اشاره «این» / «آن» + نوع، همه + اسم (N) + حرف ربط «که»

مثال‌ها:

- این همه نابسامانی (N) اجتماعی و نابرابری...
- آن شاخصی (N) که برای اندازه‌گیری این نوسانات.....

الگوی ۲۴. قید «هم» / «چنین» / «چنان» + اسم (N)

مثال:

- حاضرند چنین بهایی (N) پردازند

طراحی سامانه برچسب‌دهی به اجزای کلام برای متون فارسی □ ۳۰

الگوی ۲۵. «همین» / «همان» + اسم (N) + (فعل (V)) + را / که

مثال:

- این همان تلویزیونی (N) است (V) که چند وقت قبل.....

الگوی ۲۶. «چه» + اسم (N)

مثال:

- چه احساسی (N) داری؟

الگوی ۲۷. صفت برترین (ADJ-SUP) + اسم (N) + (فعل اسنادی / V-PRE) + حرف ربط «که» /

حرف اضافه (P) / «را»

مثال:

- این بیشترین (ADJ-SUP) رقمی (N) است (V-PRE) که در طول تاریخ تاسیس....

الگوی ۲۸. حرف ربط «اما» / «و» (CON) + اسم (N) + (اسم / صفت) + حرف ربط «که» / حرف

اضافه (P)

مثال:

- اما (CON) آزادی (N) از (P) نظر آمارتیا معنایی....

الگوی ۲۹. الف. درصد / میزان / نرخ / رشد / نسبت + (قید مقدار (ADV) / صفت (ADJ)) + اسم (N)

طراحی سامانه برچسب‌دهی به اجزای کلام برای متون فارسی □ ۳۱

الگوی ۲۹. ب. درصد / میزان / نرخ / رشد / نسبت + **صفت (ADJ)** + حرف اضافه (P)

مثال‌ها:

- میزان کم (ADJ) بارندگی (N)
- درصد پایینی (ADJ) از (P) افراد مسن....

الگوی ۳۰. حرف ربط «که» + (قید (ADV-NI-NQ-SIM) + **اسم (N)** + حرف اضافه (P) / «را» / حرف ربط «که»

مثال:

- قرار باشد که سازه‌ای (N) در (P) پاره‌گفتار موکد .....

الگوی ۳۱. حرف ربط «که» + **اسم (N)** + صفت (ADJ)

مثال:

- که جمعی (N) کثیر (ADJ) بودند از شعرا.....

الگوی ۳۲. حرف ربط «که» + **اسم (N)** + (صفت مفعولی (ADJ) + فعل (V)

مثال:

- که تاکسی‌درمی (N) شده (ADJ) است (V)

الگوی ۳۳. الف. حرف ربط «که» + (اسم (N) / (قید (ADV) + **صفت (ADJ)**

طراحی سامانه برچسب‌دهی به اجزای کلام برای متون فارسی □ ۳۲

الگوی ۳۳. ب. حرف ربط «که» + (اسم (N) / (قید (ADV)) + صفت (ADJ) + حرف عطف (CON) + **صفت (ADJ)**

مثال‌ها:

- که زیرویمی (N) ذاتی (ADJ) صداها....
- که واقعاً (ADV) تکراری (ADJ) و خسته‌کننده (ADJ) است

الگوی ۳۴. الف. ضمیر (PRO) + **اسم (N)**

الگوی ۳۴. ب. ضمیر (PRO) + **اسم (N)** + حرف اضافه (P) / «را»

مثال‌ها:

- او (PRO) شهری (N) تازه را به تصرف....
- گفتن‌ها و فهمیدن‌های او (PRO) جانشینی (N) برای (P) ترس از عمل اوست

الگوی ۳۵. قید «تنها» / «فقط» (ADV-NI-NQ-SIM) + **اسم (N)** + ((گروه حرف اضافه‌ای (PP)) + فعل (V) / (فعل (V) + حرف ربط «که»

مثال:

- وزن طبیعی تنها (ADV-NI-NQ-SIM) وقتی (N) به دست (PP) می‌آید (V) که....

الگوی ۳۶. زبان / لهجه + **صفت (ADJ)**

مثال:

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

### ۳-۳-۳- شرح روش انجام پژوهش

به منظور دستیابی به هدف این بخش از 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 شده است که مبنای آنها دو معیار کلی «صحت»<sup>۲۴</sup> و «معیار اف»<sup>۲۵</sup> است. پیش از معرفی این دو معیار و ارزیابی نرم‌افزار رفع ابهام کننده از برچسب نحوی هم‌نگاره‌های اسمی و صفتی فارسی مختوم به «ی» لازم است متنی به عنوان ورودی (به عنوان نمونه) به ابزار هضم وارد شود و برچسب‌گذاری انجام شود تا بتوان در مراحل بعدی عملکرد ابزار را با در نظر گرفتن الگوهای رفع ابهام کننده از برچسب‌ها و بدون در نظر گرفتن الگوها مورد ارزیابی قرار داد. این متن را به عنوان نمونه در اینجا ذکر می‌کنیم و این متن برای ارزیابی استفاده نشده است؛ متن استفاده شده برای ارزیابی یک‌دهم پیکره بی‌جن‌خان بوده است.

متن ورودی به صورت زیر است:

اهالی موسیقی این‌بار در پیگیری مطالبات‌شان از دولت و وزارت ارشاد جدی‌تر از همیشه‌اند؛ مطالباتی که بخش عمده‌ای از آنها درباره لغو کنسرت‌هایی است که مجوز اجرای آنها را وزارت ارشاد صادر کرده است. این حاشیه‌ای بود که در طول دو سال گذشته فراز و فرودهای بسیاری را تجربه کرد و تا آنجا پیش رفت که به استعفای علی جنتی، وزیر پیشین ارشاد انجامید. هرچند لغو کنسرت‌های موسیقی در شهرستان‌ها مساله‌ای نبود که با انتخاب سیدرضا صالحی‌امیری به عنوان وزیر ارشاد جدید متوقف شود. حالا در فاصله چند روز مانده به معرفی کابینه جدید و گمانه‌زنی‌هایی که برای انتخاب وزیر ارشاد دولت دوازدهم مطرح شده حوزه موسیقی باز هم با همین حاشیه‌ها دست و پنجه نرم می‌کند. آخرین آنها کنسرت شهرام و حافظ ناظری در شهر قوچان بود؛ کنسرتی که جمعه گذشته با دستور دادستانی این شهر

<sup>24</sup> Accuracy

<sup>25</sup> F-Measure

لغو شد. این‌طور شد که ناظری‌ها از خیر اجرا در قوچان گذشتند تا منتظر اجرای‌شان در روزهای میانی مردادماه در تهران و سالن وزارت کشور بمانند. اما لغو این کنسرت در چند روز گذشته واکنش‌های مختلف اهالی موسیقی را در پی داشت و کار تا جایی بالا گرفت که خانه موسیقی قول برگزاری جلسه‌ای فوق‌العاده برای رسیدگی به این موضوع را داد. اما عدم واکنش مسوولان ارشاد به لغو این کنسرت‌ها باعث شد تا خانه موسیقی تنها به برگزاری جلسه فوق‌العاده‌اش اکتفا نکند. این خانه روز گذشته اعلام کرد گردهمایی بزرگ اهالی موسیقی در اعتراض به لغو غیرقانونی کنسرت‌ها را روز چهارشنبه برگزار می‌کند.

متن فوق به صورت زیر توسط ابزار هضم برچسب‌گذاری شده است:

اهالی/Ne/موسیقی/Ne/این‌بار/AJ/در/P/پیگیری/Ne/مطالبات‌شان/N/از/P/دولت/N/و/CONJ/  
وزارت/Ne/ارشاد/N/جدی‌تر/AJ/از/P/همیشه‌اند/V/؛/PUNC/مطالباتی/N/که/CONJ/بخش/Ne/  
عمده‌ای/AJ/از/P/آنها/PRO/درباره/Pe/لغو/Ne/کنسرت‌هایی/N/است/V/که/CONJ/مجوز/Ne/  
اجرای/Ne/آنها/PRO/را/POSTP/وزارت/Ne/ارشاد/N/صادر/AJ/کرده\_است/PUNC./V/

این/DET/حاشیه‌ای/N/بود/V/که/CONJ/در/P/طول/Ne/دو/NUM/سال/Ne/گذشته/ADV/  
فراز/N/و/CONJ/فرودهای/Ne/بسیاری/AJ/را/POSTP/تجربه/N/کرد/V/و/CONJ/تا/P/  
آنجا/ADV/پیش/ADV/رفت/V/که/CONJ/به/P/استعفای/Ne/علی/Ne/جنتی/N/،/PUNC/  
وزیر/Ne/پیشین/AJ/ارشاد/N/انجامید/PUNC./V/

هرچند/CONJ/لغو/Ne/کنسرت‌های/Ne/موسیقی/N/در/P/شهرستان‌ها/N/مساله‌ای/N/نبود/V/  
که/CONJ/با/P/انتخاب/Ne/سیدرضا/N/صالحی/Ne/امیری/N/به/P/عنوان/Ne/وزیر/Ne/  
ارشاد/Ne/جدید/AJ/متوقف/AJ/شود/PUNC./V/

حالا/ADV/در/P/فاصله/Ne/چند/DET/روز/N/مانده/V/به/P/معرفی/Ne/کابینه/Ne/جدید/AJ/  
و/CONJ/گمانه‌زنی‌هایی/N/که/CONJ/برای/Pe/انتخاب/Ne/وزیر/Ne/ارشاد/Ne/دولت/Ne/  
دوازدهم/NUM/مطرح/AJ/شده/V/حوزه/Ne/موسیقی/N/باز/ADV/هم/CONJ/با/P/  
همین/DET/حاشیه‌ها/N/دست/N/و/CONJ/پنجه/N/نرم/AJ/می‌کند/PUNC./V/

آخرین/NUM آنها/PRO/کنسرت/Ne/شهرام/N و/CONJ/حافظ/Ne/ناظری/N در/P/شهر/Ne/ قوچان/N بود/V ؛/PUNC/کنسرتی/N که/CONJ/جمعه/Ne/گذشته/ADV/با/P/دستور/Ne/ دادستانی/Ne/این/DET/شهر/Ne/لغو/N شد/PUNC./V/

این‌طور/N شد/V که/CONJ/ناظری‌ها/N از/P/خیر/N اجرا/N در/P/قوچان/N گذشتند/V/ تا/CONJ/منتظر/AJe/اجرای‌شان/N در/P/روزهای/Ne/میانی/AJe/مردادماه/N در/P/تهران/N و/CONJ/سالن/Ne/وزارت/Ne/کشور/N بمانند/PUNC./V/

اما/CONJ/لغو/Ne/این/DET/کنسرت/N در/P/چند/DET/روز/Ne/گذشته/ADV/ واکنش‌های/Ne/مختلف/AJe/اهالی/Ne/موسیقی/N را/POSTP/در/P/پی/N داشت/V و/CONJ/ کار/N تا/P/جایی/N بالا/AJ/گرفت/V که/CONJ/خانه/Ne/موسیقی/N قول/Ne/برگزاری/Ne/ جلسه‌ای/N فوق‌العاده/AJ/برای/Pe/رسیدگی/N به/P/این/DET/موضوع/N را/POSTP/داد/V/ PUNC./

اما/CONJ/عدم/Ne/واکنش/Ne/مسوولان/Ne/ارشاد/N به/P/لغو/Ne/این/DET/کنسرت‌ها/N باعث/AJ/شد/V تا/CONJ/خانه/Ne/موسیقی/N تنها/ADV/به/P/برگزاری/Ne/جلسه/Ne/ فوق‌العاده‌اش/AJ/اکتفا/N نکند/PUNC./V/

این/DET/خانه/N روز/Ne/گذشته/ADV/اعلام/N کرد/V/گردهمایی/Ne/بزرگ/AJe/ اهالی/Ne/موسیقی/N در/P/اعتراض/N به/P/لغو/Ne/غیرقانونی/AJe/کنسرت‌ها/N را/POSTP/ روز/Ne/چهارشنبه/N برگزار/AJ/می‌کند/PUNC./V/

قبل از ارائه برون‌داد نمونه‌ای از اعمال الگوهای رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی فارسی مختوم به «ی»، لازم است درباره این نرم‌افزار (رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی فارسی مختوم به «ی») توضیحاتی اضافه شود:

برای پیاده‌سازی قوانین از زبان برنامه‌نویسی پایتون 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. برنامه به این صورت عمل می‌کند که با پیمایش در کلیه واژه‌های متن، هرگاه با واژه‌ای مواجه شود که به «ی» ختم می‌شود، تک‌تک قوانین را در

مورد این واژه بررسی می‌کند. به این معنا که این واژه را به عنوان واژه مختوم به «ی» در هر قاعده در نظر می‌گیرد و بررسی می‌کند که آیا با الگوی ارائه شده در قاعده همخوانی دارد یا خیر. به منظور آشنایی با نحوه عملکرد برنامه، چگونگی پیاده‌سازی یکی از قوانین برای نمونه تشریح می‌شود:

قاعده ۳۶: زبان/لهجه + صفت (ADJ)

برای یافتن نمونه‌های منطبق با این الگو بررسی می‌شود که آیا واژه قبل از واژه مختوم به «ی»، واژه «زبان» یا «لهجه» است؟ اگر اینگونه بود برچسب «صفت» به این واژه اختصاص داده می‌شود. در ادامه، ابتدا به معرفی این دو معیار ارزیابی پرداخته می‌شود، سپس در خصوص ابزار «هضم» و مسائل مربوط به استفاده از این ابزار توضیحات بیشتری ارائه می‌شود. در پایان نیز انواع ارزیابی‌های انجام شده و نتایج مربوط به آن‌ها شرح داده خواهد شد.

### ۳-۳-۱- معیارهای ارزیابی

همانطور که اشاره شد، دو معیار ارزیابی مورد استفاده در این پژوهش «صحت» و «معیار اف» هستند. در ارزیابی‌های انجام شده، «معیار صحت» در مواردی استفاده شده است که ارزیابی کلی نرم‌افزار در سطح تمامی برچسب‌ها مد نظر بوده و «معیار اف» نیز هنگامی به کار رفته است که هدف، ارزیابی جزئی دقت سامانه در سطح تک‌تک برچسب‌ها به صورت مجزا بوده است.

#### ۳-۳-۱-۱- معیار «صحت»

به ساده‌ترین شکل، این معیار 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.

برای روشن‌تر شدن موضوع می‌شود مثال زیر را مطرح کرد:

طراحی سامانهٔ برچسب‌دهی به اجزای کلام برای متون فارسی □ ۳۷

در جدول زیر، ردیف اول واژه‌های جملهٔ مورد نظر است، ردیف دوم برچسب‌هایی را نشان می‌دهد که توسط نیروی انسانی به واژه‌ها اختصاص داده شده است و در ردیف سوم، برچسب‌هایی را می‌بینیم که سیستم برچسب‌دهی اجزای کلام برای واژه‌ها مشخص کرده است.

جدول ۳-۱- مثال برای معیار صحت

| واژه         | او  | نظرات | شخصی | خود | را | به | زیبا | تحمیل | می‌کرد | .    |
|--------------|-----|-------|------|-----|----|----|------|-------|--------|------|
| برچسب انسانی | PRO | N     | ADJ  | PRO | P  | P  | N    | N     | V      | DELM |
| برچسب ماشینی | PRO | N     | N    | PRO | P  | P  | ADJ  | N     | V      | DELM |

همانطور که ملاحظه می‌شود، از مجموع ۱۰ برچسب، ۸ برچسب مشخص‌شده توسط سیستم با برچسب‌های انسانی انطباق دارند و در دو مورد (یعنی صفت «شخصی» و اسم خاص «زیبا»)، سیستم برچسب‌ها را به درستی تشخیص داده نشده است. بنابراین مقدار معیار صحت برای این مثال ۸۰ درصد خواهد بود که طبق فرمول زیر محاسبه شده است:

$$\text{صحت (Accuracy)} = \frac{\text{تعداد برچسب‌های صحیح}}{\text{تعداد کل برچسب‌ها}} = \frac{8}{10} = 80\%$$

### ۳-۳-۱-۲- معیار اف

این معیار از ترکیب دو معیار دقت<sup>۲۶</sup> و فراخوانی<sup>۲۷</sup> به دست می‌آید:

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

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

$$\text{معیار اف} = \frac{2 \times P \times R}{P + R}$$

اینجا هم به منظور روشن شدن بیشتر موضوع مثالی ذکر می‌شود:

---

<sup>26</sup> Precision

<sup>27</sup> Recall

جدول ۳-۲- مثال برای معیار اف

| واژه         | پردازش | به | چنین | مسائلی | بخشی | از | وظیفه | ما  | است | .    |
|--------------|--------|----|------|--------|------|----|-------|-----|-----|------|
| برچسب انسانی | N      | P  | SPEC | N      | N    | P  | N     | PRO | V   | DELM |
| برچسب ماشینی | N      | P  | SPEC | N      | ADJ  | P  | N     | PRO | V   | DELM |

معیارهای دقت، فراخوانی و معیار اف برای برچسب N در مثال فوق به شکل زیر محاسبه می‌شوند:

$$P = \frac{3}{3} = 1$$

$$R = \frac{3}{4} = 0.75$$

$$F1 = \frac{2 \times 1 \times 0.75}{1 + 0.75} \approx 0.85$$

همانطور که ملاحظه می‌شود در این مثال هرچند برچسب‌زن در تمام مواردی که به واژه‌ها برچسب N اختصاص داده است درست عمل کرده است (در نتیجه دقت آن صد درصد بوده است)، اما تمام مواردی که برچسب انسانی آن‌ها N بوده است را نتوانسته است درست تشخیص بدهد (فراخوانی سیستم کامل نبوده). به منظور تأثیر دادن هر دوی این معیارها، از «معیار اف» استفاده می‌شود که مقدار آن در این مثال نزدیک به ۸۵ درصد است.

همانگونه که ذکر شد، ابزار هضم یک کتابخانه پایتون<sup>۲۸</sup> است که امکانات مختلفی برای پیش‌پردازش زبان فارسی در اختیار محققین قرار می‌دهد. از جمله این امکانات عبارتند از نرمال‌سازی<sup>۲۹</sup>، واحدسازی/قطعه‌قطعه کردن متن، سرواژه‌سازی<sup>۳۰</sup>، برچسب‌زنی اجزای کلام و تجزیه نحوی<sup>۳۱</sup>.

برچسب‌زنی اجزای کلام هضم به صورت پیش‌فرض از مدلی استفاده می‌کند که با «پیکره متنی زبان فارسی» آموزش داده شده است. از سویی با توجه به اینکه در پژوهش‌های پیشین مرتبط با پژوهش حاضر نگارنده («بررسی ساخت‌واژی هم‌نگاره‌های اسمی و صفتی به منظور کمک به برچسب‌دهی «اسم» به کلیدواژه‌ها در پیکره‌های علمی» و «رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی فارسی») از «پیکره بی‌جن‌خان» برای استخراج الگوها استفاده شده است، در این بخش از پژوهش نیز به دو دلیل مهم باید از همین داده به عنوان داده حاوی برچسب‌های نیروی انسانی جهت انجام ارزیابی‌ها و بررسی تأثیر افزودن الگوها استفاده کرد:

– اول اینکه مجموعه برچسب‌های<sup>۳۲</sup> پیکره متنی زبان فارسی (و در نتیجه مجموعه برچسب‌های هضم) با مجموعه برچسب‌های پیکره بی‌جن‌خان متفاوت است بنابراین نمی‌شود مستقیماً برچسب‌زنی با هضم را با برچسب‌های پیکره بی‌جن‌خان سنجید و الگوهای استخراج‌شده را بر اساس برچسب‌زنی هضم اعمال کرد.

– ثانیاً پیکره متنی زبان فارسی، که هضم بر اساس آن آموزش داده شده است، یک پیکره حدود ۱۰ میلیون کلمه‌ای است که پیکره بی‌جن‌خان (با حجم حدود ۲,۶ میلیون کلمه) خود زیرمجموعه‌ای از آن است. این مسئله مشکل دیگری را برای ارزیابی برچسب‌زنی به وجود می‌آورد و آن اینکه اصولاً نباید داده‌ای که قبلاً برای آموزش سیستم به کار رفته است برای ارزیابی آن نیز به کار گرفته شود.

---

<sup>28</sup> Python Library

<sup>29</sup> Normalization

<sup>30</sup> Lemmatization

<sup>31</sup> Syntactic parsing

<sup>32</sup> Tagset

راه حل این است که مدل جدیدی برای برچسب‌زن اجزای کلام هضم با استفاده از داده بی‌جن‌خان آموزش داده شود. در نتیجه مجموعه برچسب‌های هضم در این حالت همان مجموعه برچسب‌های پیکره بی‌جن‌خان خواهند بود که الگوها نیز پیش از این بر اساس آن‌ها تعریف شده‌اند.

با توجه به توضیحات فوق، مطابق روال موجود در پژوهش‌های مشابه، کل پیکره بی‌جن‌خان به ۱۰ بخش تقسیم شد که ۹ بخش آن برای آموزش سیستم و یک بخش برای ارزیابی استفاده شد. به این ترتیب داده آموزش<sup>۳۳</sup> و داده آزمون<sup>۳۴</sup> کاملاً از هم مجزا هستند و هنگام ارزیابی مشکل تکراری بودن داده‌ها وجود ندارد.

نکته دیگر در خصوص ابزار هضم این است که هم در مدل پیش‌فرض برچسب‌زن اجزای کلام و هم در مدلی که بر اساس پیکره بی‌جن‌خان آموزش داده شده است، برچسب‌های درشت<sup>۳۵</sup> در نظر گرفته شده‌اند؛ به این معنا که مثلاً برای واژه‌ای چون «دست‌اندرکاران» برچسب «N» (اسم) اختصاص داده می‌شود و نه برچسب «N\_PL\_COM» (اسم. جمع. عام). با توجه به این مطلب، مجموعه برچسب‌های مورد استفاده در مدل آموزش داده شده با پیکره بی‌جن‌خان به شرح زیر است:

---

<sup>33</sup> Train data

<sup>34</sup> Test data

<sup>35</sup> Coarse-grained

طراحی سامانهٔ برچسب‌دهی به اجزای کلام برای متون فارسی □ ۴۲

جدول ۳-۳- مجموعهٔ برچسب‌های پیکرهٔ بی‌جن‌خان

| #   | برچسب | عنوان             |
|-----|-------|-------------------|
| ۱.  | ADJ   | صفت               |
| ۲.  | ADV   | فید               |
| ۳.  | AR    | عربی              |
| ۴.  | CON   | حرف ربط           |
| ۵.  | DELM  | جداکننده          |
| ۶.  | DET   | حرف تعریف         |
| ۷.  | IF    | ادات شرط          |
| ۸.  | INT   | حرف صوت           |
| ۹.  | MORP  | تکواژ             |
| ۱۰. | MQUA  | سورگستر           |
| ۱۱. | MS    | علامت ریاضی       |
| ۱۲. | N     | اسم               |
| ۱۳. | NN    | گستره عدد         |
| ۱۴. | NP    | گروه اسمی         |
| ۱۵. | OH    | حرف ندا           |
| ۱۶. | OHH   | منادی             |
| ۱۷. | P     | حرف اضافه         |
| ۱۸. | PP    | گروه حرف اضافه‌ای |
| ۱۹. | PRO   | ضمیر              |
| ۲۰. | PS    | جمله‌واره         |
| ۲۱. | QUA   | سور               |
| ۲۲. | SPEC  | کیفیت‌نما         |
| ۲۳. | V     | فعل               |

طراحی سامانهٔ برچسب‌دهی به اجزای کلام برای متون فارسی □ ۴۳

مسئلهٔ فوق مشکلاتی را برای ارزیابی الگوهایی که بر اساس برچسب‌های ریز<sup>۳۶</sup> تدوین شده بودند به وجود می‌آورد. این قواعد عبارتند از:

جدول ۳-۴- برچسب‌های مشکل ساز در الگوها

| شمارهٔ الگو | برچسب(های) مشکل ساز | عنوان برچسب              |
|-------------|---------------------|--------------------------|
| ۳           | ADJ_SUP             | صفت عالی                 |
| ۱۹، ۱۸، ۶   | N_SING_CN           | اسم مفرد عدد اصلی        |
| ۲۷          | V_PRE، ADJ_SUP      | صفت عالی، فعل اسنادی     |
| ۳۰          | ADV_NI_NQ_SIM       | قید غیرپرشی غیر کمی ساده |

به همین دلیل چهار برچسب ADJ\_SUP، N\_SING\_CN، V\_PRE و ADV\_NI\_NQ\_SIM نیز به مجموعهٔ برچسب‌های مورد استفاده اضافه شد و در نتیجه مجموعهٔ برچسب نهایی شامل ۲۷ برچسب است.

### ۳-۳-۲- آزمایش‌ها و نتایج

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 که در تمام آن‌ها برچسب‌زن هضم با دادهٔ آموزش (۰٫۹ بی‌جن‌خان) آموزش داده شده و برچسب‌زنی بر روی دادهٔ آزمون (۰٫۱ بی‌جن‌خان) انجام و برچسب‌های هضم با برچسب‌های انسانی دادهٔ آزمون مقایسه شده‌اند. آزمایش‌های انجام شده به شرح زیر هستند:

<sup>36</sup> Fine-grained

### ۳-۳-۲-۱- دقت کلی برچسب‌زن اجزای کلام هضم بدون الگوها و با الگوها

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 شد، این بار با در نظر گرفتن الگوها مجدداً برچسب‌ها مورد بازبینی قرار گرفت و به صورت زیر نمایش داده می‌شود:

متن نمونه:

اهالی موسیقی این‌بار در پیگیری مطالبات‌شان از دولت و وزارت ارشاد جدی‌تر از همیشه‌اند؛ مطالباتی که بخش عمده‌ای از آنها درباره لغو کنسرت‌هایی است که مجوز اجرای آنها را وزارت ارشاد صادر کرده است. این حاشیه‌ای بود که در طول دو سال گذشته فراز و فرودهای بسیاری را تجربه کرد و تا آنجا پیش رفت که به استعفای علی جنتی، وزیر پیشین ارشاد انجامید. هرچند لغو کنسرت‌های موسیقی در شهرستان‌ها مساله‌ای نبود که با انتخاب سیدرضا صالحی‌امیری به عنوان وزیر ارشاد جدید متوقف شود. حالا در فاصله چند روز مانده به معرفی کابینه جدید و گمانه‌زنی‌هایی که برای انتخاب وزیر ارشاد دولت دوازدهم مطرح شده حوزه موسیقی باز هم با همین حاشیه‌ها دست و پنجه نرم می‌کند. آخرین آنها کنسرت شهرام و حافظ ناظری در شهر قوچان بود؛ کنسرتی که جمعه گذشته با دستور دادستانی این شهر لغو شد. این‌طور شد که ناظری‌ها از خیر اجرا در قوچان گذشتند تا منتظر اجرای‌شان در روزهای میانی مردادماه در تهران و سالن وزارت کشور بمانند. اما لغو این کنسرت در چند روز گذشته واکنش‌های مختلف اهالی موسیقی را در پی داشت و کار تا جایی بالا گرفت که خانه موسیقی قول برگزاری جلسه‌ای فوق‌العاده برای رسیدگی به این موضوع را داد. اما عدم واکنش مسوولان ارشاد به لغو این کنسرت‌ها

باعث شد تا خانه موسیقی تنها به برگزاری جلسه فوق‌العاده‌اش اکتفا نکند. این خانه روز گذشته اعلام کرد  
گردهمایی بزرگ اهالی موسیقی در اعتراض به لغو غیرقانونی کنسرت‌ها را روز چهارشنبه برگزار می‌کند.

|   |         |       |             |
|---|---------|-------|-------------|
| - | -       | N     | اهالی       |
| - | -       | N     | موسیقی      |
| - | -       | ADV   | این‌بار     |
| - | -       | P     | در          |
| N | rule01a | N     | پیگیری      |
| - | -       | N     | مطالبات‌شان |
| - | -       | P     | از          |
| - | -       | N     | دولت        |
| - | -       | CON   | و           |
| - | -       | N     | وزارت       |
| - | -       | N     | ارشاد       |
| - | -       | ADJ   | جدی‌تر      |
| - | -       | P     | از          |
| - | -       | V-PRE | همیشه‌اند   |
| - | -       | DELM  | ؛           |
| - | -       | N     | مطالباتی    |
| - | -       | CON   | که          |
| - | -       | N     | بخش         |
| - | -       | ADJ   | عمده‌ای     |
| - | -       | P     | از          |
| - | -       | PRO   | آنها        |
| - | -       | P     | درباره      |

|   |         |           |            |
|---|---------|-----------|------------|
| - | -       | N         | لغو        |
| - | -       | N         | کنسرت‌هایی |
| - | -       | V-PRE     | است        |
| - | -       | CON       | که         |
| - | -       | N         | مجوز       |
| - | -       | N         | اجرای      |
| - | -       | PRO       | آنها       |
| - | -       | P         | را         |
| - | -       | N         | وزارت      |
| - | -       | N         | ارشاد      |
| - | -       | ADJ       | صادر       |
| - | -       | ADJ       | کرده       |
| - | -       | V-PRE     | است        |
| - | -       | DELM      | .          |
| - | -       | DET       | این        |
| N | rule23a | N         | حاشیه‌ای   |
| - | -       | V         | بود        |
| - | -       | CON       | که         |
| - | -       | P         | در         |
| - | -       | N         | طول        |
| - | -       | N-SING-CN | دو         |
| - | -       | N         | سال        |
| - | -       | ADJ       | گذشته      |
| - | -       | N         | فراز       |
| - | -       | CON       | و          |
| - | -       | N         | فرودهای    |

|   |         |      |            |
|---|---------|------|------------|
| - | -       | ADJ  | بسیاری     |
| - | -       | P    | را         |
| - | -       | N    | تجربه      |
| - | -       | V    | کرد        |
| - | -       | CON  | و          |
| - | -       | P    | تا         |
| - | -       | N    | آنجا       |
| - | -       | N    | پیش        |
| - | -       | V    | رفت        |
| - | -       | CON  | که         |
| - | -       | P    | به         |
| N | rule01a | N    | استعفای    |
| - | -       | N    | علی        |
| N | rule22a | N    | جنتی       |
| - | -       | DELM | ‘          |
| - | -       | N    | وزیر       |
| - | -       | ADJ  | پیشین      |
| - | -       | N    | ارشاد      |
| - | -       | V    | انجامید    |
| - | -       | DELM | .          |
| - | -       | CON  | هر چند     |
| - | -       | N    | لغو        |
| - | -       | N    | کنسرت‌های  |
| - | -       | N    | موسیقی     |
| - | -       | P    | در         |
| - | -       | N    | شهرستان‌ها |

|   |         |           |          |
|---|---------|-----------|----------|
| - | -       | N         | مساله‌ای |
| - | -       | V         | نبود     |
| - | -       | CON       | که       |
| - | -       | P         | با       |
| - | -       | N         | انتخاب   |
| - | -       | N         | سیدرضا   |
| - | -       | N         | صالحی    |
| - | -       | N         | امیری    |
| - | -       | P         | به       |
| - | -       | N         | عنوان    |
| - | -       | N         | وزیر     |
| - | -       | N         | ارشاد    |
| - | -       | ADJ       | جدید     |
| - | -       | ADJ       | متوقف    |
| - | -       | V         | شود      |
| - | -       | DELM      | .        |
| - | -       | ADV       | حالا     |
| - | -       | P         | در       |
| - | -       | N         | فاصله    |
| - | -       | N-SING-CN | چند      |
| - | -       | N         | روز      |
| - | -       | ADJ       | مانده    |
| - | -       | P         | به       |
| N | rule01a | N         | معرفی    |
| - | -       | N         | کابینه   |
| - | -       | ADJ       | جدید     |

|   |   |         |                |
|---|---|---------|----------------|
| - | - | CON     | و              |
| - | - | N       | گمانه‌زنی‌هایی |
| - | - | CON     | که             |
| - | - | P       | برای           |
| - | - | N       | انتخاب         |
| - | - | N       | وزیر           |
| - | - | N       | ارشاد          |
| - | - | N       | دولت           |
| - | - | ADJ     | دوازدهم        |
| - | - | ADJ     | مطرح           |
| - | - | ADJ     | شده            |
| - | - | N       | حوزه           |
| - | - | N       | موسیقی         |
| - | - | ADJ     | باز            |
| - | - | CON     | هم             |
| - | - | P       | با             |
| - | - | DET     | همین           |
| - | - | N       | حاشیه‌ها       |
| - | - | N       | دست            |
| - | - | CON     | و              |
| - | - | N       | پنجه           |
| - | - | ADJ     | نرم            |
| - | - | V       | می‌کند         |
| - | - | DELM    | .              |
| - | - | ADJ-SUP | آخرین          |
| - | - | PRO     | آنها           |

طراحی سامانهٔ برچسب‌دهی به اجزای کلام برای متون فارسی □ ۵۰

|   |   |      |          |
|---|---|------|----------|
| - | - | N    | کنسرت    |
| - | - | N    | شهرام    |
| - | - | CON  | و        |
| - | - | N    | حافظ     |
| - | - | N    | ناظری    |
| - | - | P    | در       |
| - | - | N    | شهر      |
| - | - | N    | قوچان    |
| - | - | V    | بود      |
| - | - | DELM | ؛        |
| - | - | N    | کنسرتی   |
| - | - | CON  | که       |
| - | - | N    | جمعه     |
| - | - | ADJ  | گذشته    |
| - | - | P    | با       |
| - | - | N    | دستور    |
| - | - | N    | دادستانی |
| - | - | DET  | این      |
| - | - | N    | شهر      |
| - | - | N    | لغو      |
| - | - | V    | شد       |
| - | - | DELM | .        |
| - | - | N    | این‌طور  |
| - | - | V    | شد       |
| - | - | CON  | که       |
| - | - | N    | ناظری‌ها |

|   |   |      |           |
|---|---|------|-----------|
| - | - | P    | از        |
| - | - | N    | خیر       |
| - | - | N    | اجرا      |
| - | - | P    | در        |
| - | - | N    | قوچان     |
| - | - | V    | گذشتند    |
| - | - | P    | تا        |
| - | - | ADJ  | منتظر     |
| - | - | N    | اجرای‌شان |
| - | - | P    | در        |
| - | - | N    | روزهای    |
| - | - | ADJ  | میانی     |
| - | - | N    | مردادماه  |
| - | - | P    | در        |
| - | - | N    | تهران     |
| - | - | CON  | و         |
| - | - | N    | سالن      |
| - | - | N    | وزارت     |
| - | - | N    | کشور      |
| - | - | V    | بمانند    |
| - | - | DELM | .         |
| - | - | CON  | اما       |
| - | - | N    | لغو       |
| - | - | DET  | این       |
| - | - | N    | کنسرت     |
| - | - | P    | در        |

|     |         |           |            |
|-----|---------|-----------|------------|
| -   | -       | N-SING-CN | چند        |
| -   | -       | N         | روز        |
| -   | -       | ADJ       | گذشته      |
| -   | -       | N         | واکنش‌های  |
| -   | -       | ADJ       | مختلف      |
| ADJ | rule14  | N         | اهالی      |
| -   | -       | N         | موسیقی     |
| -   | -       | P         | را         |
| -   | -       | P         | در         |
| N   | rule01a | N         | پی         |
| -   | -       | V         | داشت       |
| -   | -       | CON       | و          |
| -   | -       | N         | کار        |
| -   | -       | P         | تا         |
| N   | rule01a | N         | جایی       |
| -   | -       | ADJ       | بالا       |
| -   | -       | V         | گرفت       |
| -   | -       | CON       | که         |
| -   | -       | N         | خانه       |
| -   | -       | N         | موسیقی     |
| -   | -       | N         | قول        |
| -   | -       | N         | برگزاری    |
| -   | -       | N         | جلسه‌ای    |
| -   | -       | ADJ       | فوق‌العاده |
| ADJ | rule14  | P         | برای       |
| N   | rule01a | N         | رسیدگی     |

|   |         |                   |               |
|---|---------|-------------------|---------------|
| - | -       | P                 | به            |
| - | -       | DET               | این           |
| - | -       | N                 | موضوع         |
| - | -       | P                 | را            |
| - | -       | V                 | داد           |
| - | -       | DELM              | .             |
| - | -       | CON               | اما           |
| - | -       | N                 | عدم           |
| - | -       | N                 | واکنش         |
| - | -       | N                 | مسوولان       |
| - | -       | N                 | ارشاد         |
| - | -       | P                 | به            |
| - | -       | N                 | لغو           |
| - | -       | DET               | این           |
| - | -       | N                 | کنسرت‌ها      |
| - | -       | N                 | باعث          |
| - | -       | V                 | شد            |
| - | -       | P                 | تا            |
| - | -       | N                 | خانه          |
| - | -       | N                 | موسیقی        |
| - | -       | ADV-NI-<br>NQ-SIM | تنها          |
| - | -       | P                 | به            |
| N | rule01a | N                 | برگزاری       |
| - | -       | N                 | جلسه          |
| - | -       | N                 | فوق‌العاده‌اش |
| - | -       | N                 | اکتفا         |

|     |        |      |           |
|-----|--------|------|-----------|
| -   | -      | V    | نکند      |
| -   | -      | DELM | .         |
| -   | -      | DET  | این       |
| -   | -      | N    | خانه      |
| -   | -      | N    | روز       |
| -   | -      | ADJ  | گذشته     |
| -   | -      | N    | اعلام     |
| -   | -      | V    | کرد       |
| -   | -      | N    | گردهمایی  |
| -   | -      | ADJ  | بزرگ      |
| ADJ | rule14 | N    | اهالی     |
| -   | -      | N    | موسیقی    |
| -   | -      | P    | در        |
| -   | -      | N    | اعتراض    |
| -   | -      | P    | به        |
| -   | -      | N    | لغو       |
| -   | -      | ADJ  | غیرقانونی |
| -   | -      | N    | کنسرت‌ها  |
| -   | -      | P    | را        |
| -   | -      | N    | روز       |
| -   | -      | N    | چهارشنبه  |
| -   | -      | ADJ  | برگزار    |
| -   | -      | V    | می‌کند    |
| -   | -      | DELM | .         |

نکته: ستون دوم برچسب هضم و ستون چهارم برچسب اعمال الگوها است. به منظور ارزیابی الگوها باید هضم مجدداً آموزش داده می‌شد، به همین خاطر برچسب‌های هضم در فایل مربوط به این پژوهش ممکن است با برچسب‌های هضم که از سایت هضم به دست آمده است کمی متفاوت باشد. به عنوان مثال کلمه «کنسرت» در عبارت «کنسرت شهرام ناظری»، در سایت هضم به صورت Ne برچسب گذاری شده است، زیرا کسره اضافه را با e نشان داده شده است.

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 الگوهای» و «بدون الگوها» انجام شده است. در حالت «با الگوها» تمامی الگوهای استخراج شده در طرح پژوهشی «رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی فارسی (علایی: ۱۳۹۵)» در هنگام برچسب‌زنی به کار گرفته می‌شوند. نتایج این آزمایش به شرح زیر است.

جدول ۳-۵- درصد صحت (Accuracy) کلی برچسب‌زن هضم

| بدون الگوها | با الگوها |
|-------------|-----------|
| ۹۵,۶۹۰      | ۹۴,۳۵۱    |

همانطور که ملاحظه می‌شود به کارگیری الگوهای استخراج شده موجب کاهش ۱,۳۳۹ درصدی صحت برچسب‌زنی شده است. علت اختلاف کم میان دو حالت «بدون الگوها» و «با الگوها» دامنه محدود اثرگذاری الگوها است. به عبارت دیگر قاعده‌های استخراج شده تنها در مورد واژه‌های مختوم به «ی» در داده‌آزمون، 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 داده‌آزمون را تشکیل می‌دهند و در نتیجه حداکثر میزان تأثیر الگوها می‌تواند همین مقدار باشد.

### ۳-۳-۲- دقت تک تک برچسب‌های هضم بدون الگوها و با الگوها

جدول زیر تأثیر به کارگیری الگوها بر دقت برچسب‌زنی در سطح تک تک برچسب‌ها را بر اساس معیار «اف» گزارش می‌کند:

جدول ۳-۶- دقت تک تک برچسب‌های هضم بر اساس معیار اف

| برچسب         | بدون الگوها | با الگوها | میزان تأثیر |
|---------------|-------------|-----------|-------------|
| V-PRE         | 0.94588     | 0.94576   | -0.00012    |
| N-SING-CN     | 0.67365     | 0.67319   | -0.00047    |
| V             | 0.97373     | 0.97313   | -0.00060    |
| CON           | 0.97998     | 0.97692   | -0.00306    |
| SPEC          | 0.66258     | 0.6563    | -0.00628    |
| P             | 0.98828     | 0.97823   | -0.01005    |
| ADV           | 0.85263     | 0.84094   | -0.01170    |
| N             | 0.97103     | 0.95683   | -0.01421    |
| ADJ           | 0.92429     | 0.88862   | -0.03567    |
| ADV-NI-NQ-SIM | 0.69769     | 0.66057   | -0.03713    |
| PRO           | 0.96231     | 0.89667   | -0.06565    |
| QUA           | 0.83837     | 0.73856   | -0.09981    |
| PS            | 0.58065     | 0.42857   | -0.15207    |
| MORP          | 0.56637     | 0.38000   | -0.18637    |
| INT           | 0.43750     | 0.07692   | -0.36058    |
| OH            | 0.84615     | 0.33333   | -0.51282    |

همانطور که ملاحظه می‌شود به کارگیری الگوها در مورد هیچ یک از برچسب‌ها تأثیر مثبت نداشته است. برچسب‌هایی که میزان تأثیر افزودن الگوها بر تشخیص آن‌ها صفر بوده است در جدول فوق گزارش نشده‌اند.

### ۳-۳-۳-۳- تأثیر هر قاعده در بهبود دقت برچسب‌زن

هدف از این آزمایش بررسی تأثیر تک‌تک الگوها بر بهبود دقت برچسب‌زن بوده است. جدول زیر نتیجهٔ این آزمایش را نشان می‌دهد. لازم به توضیح است که برای محاسبهٔ تأثیر هر قاعده، معیار صحت بر اساس تمام مواردی که قاعده اعمال شده است یک بار بدون الگوها و یک بار با الگوها محاسبه شده است. الگوهایی که در جدول زیر ذکر نشده‌اند در مورد هیچ یک از واژه‌های پیکرهٔ آزمون قابل اعمال نبوده‌اند.

جدول ۳-۷- میزان تأثیر تک‌تک الگوها

| قاعده           | بدون الگوها | با الگوها | میزان تأثیر |
|-----------------|-------------|-----------|-------------|
| rule01a         | 0.97659     | 0.87992   | -0.09668    |
| rule01b         | 0.97076     | 0.83041   | -0.14035    |
| rule02a         | 0.87500     | 0.63333   | -0.24167    |
| rule02b         | 1.00000     | 0.83333   | -0.16667    |
| rule03          | 0.90000     | 0.42500   | -0.47500    |
| rule05a         | 0.83784     | 0.40541   | -0.43243    |
| rule06a         | 1.00000     | 1.00000   | 0.00000     |
| rule06b         | 0.75000     | 1.00000   | +0.25000    |
| rule07          | 0.88235     | 0.43529   | -0.44706    |
| rule08          | 0.93939     | 1.00000   | +0.06060    |
| rule08DELM_hazm | 0.95745     | 0.72340   | -0.23404    |
| rule09a         | 0.94650     | 0.93631   | -0.01019    |
| rule09b         | 0.87146     | 0.76113   | -0.11032    |
| rule10a         | 0.93043     | 0.73913   | -0.19130    |
| rule10b         | 0.95475     | 0.77376   | -0.18100    |
| rule11          | 0.96335     | 0.93194   | -0.03141    |
| rule12a         | 0.94737     | 0.10526   | -0.84211    |
| rule12b         | 1.00000     | 1.00000   | 0.00000     |
| rule13          | 0.91667     | 0.83333   | -0.08333    |
| rule14          | 0.94541     | 0.45472   | -0.49069    |
| rule15a         | 0.95578     | 0.57331   | -0.38247    |
| rule15b         | 0.94521     | 0.80822   | -0.13699    |
| rule18          | 0.86900     | 0.44541   | -0.42358    |
| rule20          | 0.95604     | 0.68132   | -0.27473    |
| rule21a         | 0.94406     | 0.81119   | -0.13287    |
| rule22a         | 0.94820     | 0.57320   | -0.37500    |
| rule23a         | 0.96482     | 0.81407   | -0.15075    |

|          |         |         |         |
|----------|---------|---------|---------|
| -0.09333 | 0.82667 | 0.92000 | rule24  |
| 0.00000  | 1.00000 | 1.00000 | rule25  |
| -0.04348 | 0.93478 | 0.97826 | rule26  |
| 0.00000  | 0.95833 | 0.95833 | rule27  |
| -0.14554 | 0.77934 | 0.92488 | rule28  |
| -0.38095 | 0.56190 | 0.94286 | rule29a |
| -0.52727 | 0.45455 | 0.98182 | rule30  |
| -0.18000 | 0.82000 | 1.00000 | rule31  |
| -0.26316 | 0.52632 | 0.78947 | rule32  |
| -0.50433 | 0.43723 | 0.94156 | rule33a |
| -0.33953 | 0.60930 | 0.94884 | rule34a |
| 0.00000  | 1.00000 | 1.00000 | rule35  |
| -0.19231 | 0.80769 | 1.00000 | rule36  |

دقت بالای ۸۰ درصد برای بسیاری از الگوها نشان‌دهندهٔ عملکرد درست آنها در مورد اکثر واژه‌هاست. اما همانطور که ملاحظه می‌شود به دلیل اینکه برچسب‌زن هضم هم در این موارد دقت خوبی داشته است، تأثیر مثبتی در عملکرد کلی دیده نمی‌شود.

نهایتاً بر اساس جدول فوق 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.

تمامی روش‌ها و معیارهایی که ذکر شد از روش‌ها و معیارهای رایج در پژوهش‌های حوزهٔ پردازش زبان طبیعی هستند و می‌توان توضیحات مربوط به آن‌ها را در بسیاری از کتاب‌های مرجع این حوزه یافت. برای مثال در جورافسکی و مارتین ۲۰۰۹، توضیحات مربوط به معیار صحت (Accuracy) و داده‌های آزمون و آموزش در فصل ۵، و توضیحات مربوط به معیار «اف» در فصل ۱۳ ارائه شده است.

## فصل چهارم

# معرفی سامانهٔ برچسب‌دهی به اجزای کلام برای متون فارسی

طراحی سامانه برچسب‌دهی به اجزای کلام برای متون فارسی □ ۶۰

## معرفی سامانه برچسب‌دهی به اجزای کلام برای متون فارسی

لینک سامانه از طریق آدرس <http://138.201.69.71> قابل مشاهده می‌باشد. با استفاده از این آدرس صفحه‌ای باز می‌شود که باید نام کاربری و گذرواژه را در قسمت مربوطه وارد کرد:

سپس صفحه زیر باز می‌شود که در قسمت مربوطه متن ورودی را می‌توان تایپ کرد و یا با فرمت txt و docx کپی و پیست نمود:



وزارت علوم، تحقیقات و فناوری  
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

دریاره سامانه | تماس یا ما

لطفاً متن خود را انتخاب کنید (با فرمت txt یا docx) و یا در کادر پایین وارد نمایید و سپس دکمه ثبت را بزنید.

بارگذاری فایل

تعداد صفحات: ۰ (هر صفحه: ۴۰۰ کلمه)

رفع ابهام

ثبت

همه حقوق از آن پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است. © ۱۳۹۶  
تهران - خیابان انقلاب اسلامی - چهارراه فلسفین - شماره ۱۰۹۰ - کد پستی ۱۳۱۵۷۷۳۳۱۴ - صندوق پستی ۱۳۱۸۵-۱۳۷۱۰ - تلفن‌خانه ۶۶۴۹۴۹۸۰ - ۶۶۴۹۵۱۴۳۰ (۰۲۱) - دورنگار ۶۶۴۶۲۲۵۴ (۰۲۱)

(۰۲۱)

همانگونه که مشخص است، در گوشه سمت راست تصویر، گزینه‌ای تحت عنوان «رفع ابهام» وجود دارد، با انتخاب این گزینه، سامانه علاوه بر برچسب‌گذاری متن، برخی از قواعد مفید رفع ابهام از هم-نگاره‌های اسمی و صفتی فارسی مختوم به «ی» را به کار می‌گیرد و در صورت اشتباه بودن برچسب نحوی، برچسب درست را به این هم‌نگاره‌ها با استفاده از قواعد مذکور، اختصاص می‌دهد.

محدودیت خاصی برای ورود متن در نظر گرفته نشده است، حتی اگر متن حاوی اشکال و جدول باشد، سامانه تنها متن موجود در آن‌ها را برای برچسب‌گذاری لحاظ می‌کند. برای نمونه، متن زیر در این قسمت به سامانه وارد شده است:

فضا و محیط کاری که در آن مشغول به فعالیت هستیم به اندازه ی تیم کاری مهم است. چون فضای کاری جایی است که در آن روزانه ساعات طولانی را مشغول به فعالیت هستیم و انجام وظایف و مسئولیت ها هستیم. بنابراین علاوه بر داشتن یک محیط کاری با تجهیزات کافی و امکانات خوب دوستانه بودن این محیط نیز مهم است که یک تیم کاری بتواند به خوبی به انجام مسئولیت هایش بپردازد.

طراحی سامانه برچسب‌دهی به اجزای کلام برای متون فارسی □ ۶۲



وزارت علوم، تحقیقات و فناوری  
پروژه، شبکه علوم و فناوری اطلاعات ایران «ایرانداک»



درباره سامانه



تماس با ما



بارگذاری فایل

لطفاً متن خود را انتخاب کنید (با فرمت docx یا txt) و یا در کادر پایین وارد نمایید و سپس دکمه ثبت را بزنید.

فضا و محیط کاری که در آن مشغول به فعالیت هستیم به اندازه ی تیم کاری مهم است. چون فضای کاری جایی است که در آن روزانه ساعات طولانی را مشغول به فعالیت هستیم و انجام وظایف و مسئولیت ها هستیم بنابراین علاوه بر داشتن یک محیط کاری با تجهیزات کافی و امکانات خوب دوستانه بودن این محیط نیز مهم است که یک تیم کاری بتواند به خوبی به انجام مسئولیت هایش بپردازد.

تعداد صفحات: ۱ (هر صفحه: ۴۰۰ کلمه)

رفع ابهام

ثبت

همه حقوق از آن پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است. © ۱۳۹۶

تهران - خیابان انقلاب اسلامی - چهارراه فلسطین - شماره ۱۰۹۰ - کد پستی ۱۳۱۵۷۷۳۳۱۴ - صندوق پستی ۱۳۱۸۵-۱۳۷۱ - تلفن خانه ۶۶۴۹۴۹۸۰ - ۶۶۹۵۱۴۳۰ (۰۲۱) - دورنگار ۶۶۴۶۲۲۵۴ (۰۲۱)

(۰۲۱)

پس از وارد کردن یا بارگذاری فایل مربوطه، گزینه « ثبت » را انتخاب می کنیم و صفحه زیر باز می شود:

طراحی سامانه برچسب‌دهی به اجزای کلام برای متون فارسی □ ۶۳



وزارت علوم، تحقیقات و فناوری  
پروژه‌های علم و فناوری اطلاعات ایران (پروژاک)

دریاره سامانه

تفاس یا ما

است

فضا و محیط کاری که در آن مشغول به فعالیت هستیم به اندازه‌ی تیم کاری مهم است

چون فضای کاری جایی است که در آن روزانه ساعات طولانی را مشغول به فعالیت هستیم و انجام وظایف و مسئولیت‌ها هستیم بنابراین علاوه بر داشتن یک محیط کاری یا تجهیزات کافی و امکانات خوب دوستانه بودن این محیط نیز مهم است که یک تیم کاری بتواند به خوبی به انجام مسئولیت‌هایش بپردازد

برچسب‌های POS

فراوانی واژه‌ها

فراوانی POS

راهنمای POS

|                   |            |
|-------------------|------------|
| صفت عالی          | صفت        |
| قید ساده          | قید        |
| حرف ربط           | عربی       |
| حرف تعریف         | جدگنده     |
| حرف صفت           | ادات شرط   |
| سور               | تکواژ      |
| علت ریاضی         | سورگستر    |
| گستره عدد         | اسم        |
| عدد               | گروه اسمی  |
| متادی             | حرف ندا    |
| گروه حرف اضافه‌ای | حرف اضافه  |
| جمله‌واره         | ضمیر       |
| فعل               | کیفیت‌نما  |
|                   | فعل استادی |

دریافت خروجی

ثبت متن جدید

همه حقوق از آن پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است. © ۱۳۹۶

تهران - خیابان انقلاب اسلامی - چهارراه فلسطین - شماره ۱۰۹۰ - کد پستی ۱۳۱۵۷۷۳۳۱۴ - صندوق پستی ۱۳۷۱-۱۳۱۸۵ - تلفن‌خانه ۶۶۴۹۴۹۸۰ - ۶۶۹۵۱۴۳۰ (۰۲۱) - دورنگار ۶۶۴۶۲۲۵۴ (۰۲۱)

(۰۲۱)

در این قسمت، همانگونه که ملاحظه می‌شود، کلمات موجود در متن برچسب‌گذاری شده‌اند و هر برچسب به رنگ مشخصی که در «راهنمای POS» موجود است، مشخص شده است. با کلیک کردن روی گزینه «فراوانی واژه‌ها»، در قسمت گوشه سمت راست، دومین گزینه، می‌توان فراوانی واژه‌ها همراه با برچسب مربوطه ملاحظه نمود:

طراحی سامانه برچسب‌دهی به اجزای کلام برای متون فارسی □ ۶۴



وزارت علوم، تحقیقات و فناوری  
پروژه‌های علوم و فناوری اطلاعات ایران (ایرانداک)

دریافت سامانه

تماس با ما

برچسب‌های POS

فرآوانی واژه‌ها

فرآوانی POS

راهنمای POS

صفت

صفت عالی

قید

قید ساده

عربی

حرف ربط

جداکننده

حرف تعریف

ادوات شرط

حرف صفت

تکواژ

سور

سورگستر

علاقه ریاضی

اسم

گستره عدد

گروه اسمی

عدد

حرف ندا

حرف اضافه

گروه حرف اضافه‌ای

ضمیر

جمله‌واره

کیفیت‌نما

فعل

فعل استادی

| ردیف | واژه  | POS        | تکرار |
|------|-------|------------|-------|
| ۱    | به    | حرف اضافه  | ۵     |
| ۲    | و     | حرف ربط    | ۴     |
| ۳    | است   | فعل استادی | ۳     |
| ۴    | محیط  | اسم        | ۳     |
| ۵    | هستیم | فعل        | ۳     |
| ۶    | کاری  | اسم        | ۳     |
| ۷    | که    | حرف ربط    | ۳     |
| ۸    | .     | جداکننده   | ۲     |
| ۹    | آن    | ضمیر       | ۲     |
| ۱۰   | انجام | اسم        | ۲     |
| ۱۱   | تیم   | اسم        | ۲     |

فقط اسم‌ها

ثبت متن جدید

دریافت خروجی

همه حقوق از آن پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است. © ۱۳۹۶

تهران - خیابان انقلاب اسلامی - چهارراه فلسطین - شماره ۱۰۹۰ - کد پستی ۱۳۱۵۷۷۳۳۱۴ - صندوق پستی ۱۳۱۸۵-۱۳۷۱ - تلفن‌خانه ۶۶۴۹۴۹۸۰ - ۶۶۹۵۱۴۳۰ (۰۲۱) - دورنگار ۶۶۴۶۲۲۵۴ (۰۲۱)

در این قسمت می‌توان روی گزینه «دریافت خروجی»، کلیک کرد و همین فهرست را در فایل اکسل مشاهده نمود و ذخیره کرد.

طراحی سامانه برچسب‌دهی به اجزای کلام برای متون فارسی □ ۶۵

همانگونه که ملاحظه می‌شود در قسمت پایین، گزینه‌ای تحت عنوان «فقط اسم‌ها» وجود دارد، با کلیک کردن روی این گزینه، در همین صفحه تنها اسم‌ها نمایش داده می‌شود (کاربرد تهیه فهرستی از اسم‌ها همراه با فراوانی آن‌ها در متون، در قسمت استخراج کلیدواژه‌ها و سامان‌دهی و طبقه‌بندی اطلاعات است).

ردیف

واژه

POS

تکرار

پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

دریافت خروجی

ثبت متن جدید

فقط اسم‌ها

برچسب‌های POS

فراوانی واژه‌ها

فراوانی POS

راهم‌های POS

صفت

صفت عالی

قید

قید ساده

عربی

حرف ربط

جداکننده

حرف تعریف

ادوات شرط

حرف صوت

تکواژ

سور

سور گستر

علاقه ریاضی

اسم

گستره عدد

گروه اسمی

عدد

حرف ندا

منادی

حرف اضافه

گروه حرف اضافه‌ای

ضمیر

جمله‌واره

فعل

کیفیت‌ما

فعل استادی

همه حقوق از آن پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است. © ۱۳۹۶

تهران - خیابان انقلاب اسلفی - چهارراه فلسطین - شماره ۱۰۹۰ - کد پستی ۱۳۱۵۷۷۳۳۱۴ - صندوق پستی ۱۳۱۸۵-۱۳۷۱ - تلفن‌خانه ۶۶۴۹۴۹۸۰-۶۶۴۹۵۱۴۳۰ (۰۲۱) - دورنگار ۶۶۴۶۲۲۵۴ (۰۲۱)

۰۲۱

طراحی سامانهٔ برچسب‌دهی به اجزای کلام برای متون فارسی □ ۶۶

با کلیک کردن روی گزینهٔ «دریافت خروجی»، می‌توان خروجی فهرست اسم‌ها و فراوانی آن‌ها را در فایل اکسل مشاهده و ذخیره نمود.

برای وارد کردن متن جدید، در هر مرحله می‌توان گزینهٔ «ثبت متن جدید» را انتخاب کرد و فرآیند را مجدداً شروع کرد.

## فصل پنجم

## نتیجه‌گیری

## 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 ساخت‌واژه 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، اسم معنی یا اشیا، تصغیر و تحییب، اسم مصدر یا حاصل مصدر)، شناسه 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. 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 پیشینه پژوهش و ذکر پژوهش‌های انجام شده در حوزه 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 تهیه سامانه 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 سامانه تهیه

طراحی سامانهٔ برچسب‌دهی به اجزای کلام برای متون فارسی □ ۶۹

شده (مبتنی بر سیستم برچسب‌دهی موجود به نام «هضم» ) جهت برچسب‌دهی خودکار اجزای کلام،  
پرداخته شده است.

## فهرست منابع فارسی

- صدقی، ف. ۱۳۹۲. رفع ابهام از هم‌نویسه‌ها در متون فارسی با روش‌های نیمه-نظارتی. دانشگاه الزهراء - دانشکده فنی.
- علایی، ا. ۱۳۹۵. رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی فارسی. پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک).
- علایی، ا. ۱۳۹۵. بررسی ساخت‌وازی هم‌نگاره‌های اسمی و صفتی به منظور کمک به برچسب‌دهی «اسم» به کلیدواژه‌ها در پیکره‌های علمی. پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک).
- محسنی، مهدی (۱۳۸۷). سیستم برچسب‌گذاری و ابهام‌زدایی خودکار اجزای کلام برای پیکره متنی زبان فارسی. دانشگاه علم و صنعت. دانشکده مهندسی کامپیوتر
- محسنی، م.، و بهروز مینایی بیدگلی (۱۳۸۸). سیستم برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی. دو فصل‌نامه پردازش علائم و داده‌ها. شماره ۲. پیاپی ۱۲.

## فهرست منابع لاتین

- Assi, M. and Haji Abdolhosseini, M. 2000. Grammatical tagging of a Persian corpus. *International journal of corpus linguistics*. 5 (1) : 69-82.
- Binot, J.L. 1987. Fragmentation and part of speech disambiguation. *Proceeding of the third conference on European chapter of the Association for computational linguistics*. Association for computational linguistics. 284-290.
- Felipe, M.M., and Zamorano, J.P. 2000. POS disambiguation and Partial Parsing Bidirectional interaction. *LREC*
- <https://github.com/sobhe/hazm>
- 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.
- Makki, R and Homayounpour, M. 2008. Word sense disambiguation of Farsi homographs using Thesaurus and Corpus. *Advances in natural language processing*. 5221 : 315-323 of the series lecture

*notes in computer science. 6<sup>th</sup> international conference GoTAL. Sweden.*

- Marquez, L., Padro, L., and Rodriguez, H. 2000. A machine learning approach to POS tagging. *Machine learning Kluwer Academic Publishers, Boston. Manufactured in the Netherlands.* 39 (1).59-91.
- Megerdooian, Karine (2004). Developing a Persian part-of-speech tagger. *Proceedings of the 1<sup>st</sup> workshop on Persian language and computer.* Pp. 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. 3<sup>RD</sup> Italian conference on computational linguistics. CLiC.*
- Ribeiro, R., Oliveira, L., and Trancoso, I. 2002. Morphosyntactic disambiguation for TTS systems. *LREC*
- Roth, D., and zelenko, D. 1998. Part Of Speech tagging using a network of linear separators. *Proceedings of the 17<sup>th</sup> international conference on computational linguistics. Association for computational linguistics.* 2.
- 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 17<sup>th</sup> international conference on computational linguistics and the*

*36<sup>th</sup> annual meeting of the association for computational linguistics (COLING-ACL`98)*. Montreal, Canada. 2 : 1398-1402.

- 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.

## **Abstract**

The aim of the present study is to check the possibility of improving the function of a POS tagger system via POS tag disambiguation of Persian noun and adjective homographs ending in <ی>. The case study in present research is “HAZM” tagger system. The POS tag disambiguation program is based on some context-sensitive rules extracted via studying homographs in syntactic contexts using knowledge of neighboring words (homographs ending in <ی> were studied in terms of the history of 10 windows) in order to decide about the right tag of the homograph based on the structure of the sentence. General evaluation of the mentioned POS disambiguation program indicates that if some of the context-sensitive rules which play a role in better POS tagging are added to “HAZM”, the general accuracy of “HAZM” rises to %95.691 which is %1.34 higher than the state of applying all rules. Finally a website for POS tagging was designed which is based on “HAZM” and it benefits from POS tag disambiguation of Persian noun and adjective homographs ending in <ی>. In this website it is possible to see the POS tag of the words in the text, the frequency of the words with their related tags and the frequency of the nouns.

**Keywords:** noun and adjective homographs ending in <ی>, HAZM, context-sensitive rules, POS accuracy, website for POS tagging



**Iranian Research Institute  
for Information Science and Technology (Irandoct)**

# **Designing a web site for POS tagging for Persian texts**

**By:**

**Elham Alayiaboobar**

**Spring 2018**