سومین کنفرانس ملی کامپیوتر ، فناوری اطلاعات و کاربردهای هوش مصنوعی
۱۳ بهمن ۱۳۹۸–دانشگاه شهید چمران اهواز

استخراج، بررسی و مقایسه باهم آیی کلمه‌ها از متن خبرهای وب‌گاه انگلیسی رادیوی صدا و سیما

احمد یوسفان، مربی، دانشگاه کاشان، yoosofan@kashan.ac.ir
دانیال ابراهیم‌زاده، دانش‌آموخته کارشناسی، دانشگاه کاشان، danial199472@gmail.com
مسعود عباسیان، دانش‌آموخته کارشناسی، msd.abasian@gmail.com

چکیده

باهم آیی عبارتی است که برای رساندن یک مفهوم یا معنی استفاده شده و شامل بیش از یک کلمه است. در این پژوهش به روش‌های گوناگون به استخراج، بررسی و مقایسهٔ باهم‌آیی‌های کلمه‌ها و دسته‌بندی آن‌ها از روی بخشی از بایگانی خبرهای وب‌گاه صدا و سیمای جمهوری اسلامی ایران پرداختیم. پس از گرفتن پایگاه دادهٔ خبرها از وب‌گاه صدا و سیمای جمهوری اسلامی ایران، آن‌ها را پردازش کردیم و داده‌های غیرخبری را از آن مجموعه حذف کردیم. سپس برچسب‌های HTML موجود در هر خبر را اصلاح کرده و با استفاده از تابع‌های آماده موجود در زبان برنامه‌نویسی پایتون، برچسب‌های HTML اصلاح‌شده را از متن اصلی اخبار جدا کردیم. در ادامه کلمه‌های متن‌های پالایش شده را به کمک بستهٔ NLTK بر پایهٔ نقش آن‌ها در جمله، برچسب‌گذاری و ریشه‌یابی نمودیم. سپس باهم‌آیی‌های کلمه‌ها را بر پایهٔ معیارهای تناظر به دست آوردیم و بعد مقایسه و تحلیل کردیم. همچنین در این کار اثراتِ کلمه‌های ایست‌واژه موجود در متن خبرها را در استخراجِ باهم‌آیی‌ها، مورد بررسی و تحلیل قرار دادیم. در این پژوهش از میانِ روش‌های موجود در این زمینه، مهم‌ترین و کم‌خطاترین روش‌ها را برگزیده و ترکیب کردیم و به نتیجه‌های سودمندی و مفیدی برای باهم‌آیی‌های کلمه‌ها در متن خبرهای این وب‌گاه دست یافتیم.

واژه‌های کلیدی: باهم‌آیی، ایست‌واژه، معیار تناظر، نقش کلمات در جمله، NLTK.

۱- مقدمه

بررسی باهم‌آیی[1] در هر زبان از اهمیت ویژه‌ای نزد زبان‌شناسان برخوردار است. نخستین بار اصطلاح باهم‌آیی را دانشمند زبان‌شناس فرانسوی جی.آر.فرث در نظریه معنایی خود مطرح کرد. او این پدیده زبانی را معنا بنیاد فرض کرد نه دستوری، و آن را برای نامیدن و مشخص کردن ترکیبات، بر پایهٔ رابطهٔ معنایی_اصطلاحی و بسامد وقوع آنها در زبان به کار برد. به نظر او، هم‌نشینی یکی از شیوه‌های بیان معنا است [۱].

در بیانی دیگر بررسی چگونگی کنار هم قرار گرفتن کلمه‌ها نسبت به یکدیگر در متن‌های گوناگون را باهم‌آیی کلمه‌ها می‌نامند. باهم‌آیی کلمه‌ها اطلاعات بیشتری را نسبت به فراوانی کلمه‌ها در متن‌ها و دیگر بررسی‌های ساده در اختیار پژوهشگران حوزهٔ پردازش زبان و زبان‌شناسی می‌گذارد.

یافتن و مقایسهٔ باهم‌آیی‌های کلمه‌ها در متن‌ها عمری طولانی نسبت به دیگر حوزه‌های پژوهشی رایانه دارد و پژوهشگران زبان‌شناسی کار روی باهم‌آیی کلمه‌ها را پیش از پژوهشگران رایانه آغاز کردند و کتاب‌ها، مقالات، پایگاه کلمات و فرهنگ‌های گوناگون را در این زمینه نوشته یا آماده کردند. امروزه به کمک برنامه نویسی رایانه‌ای می‌توان حجم بیشتری از متن‌ها را پردازش نمود و روش‌هایی را به کار برد که پیش از این به دلیل زمانبر بودن در حالت دستی شدنی نبودند. در حالت دستی اغلب در محدودهٔ صد تا هزار سند بررسی و پردازش انجام می‌شود در حالی که به کمک رایانه پردازشِ هزاران سندِ متنی کاری متداول است. بنابراین نتیجه‌های به‌دست آمده به دلیل بررسی خودکار تعداد متن‌های زیادتر جامعیت بیشتری دارد. دقت روش‌های رایانه‌ای به برنامه نویسی و دقت به جنبه‌های گوناگون متن وابسته است. یافتن آغاز و پایان کلمه‌ها و جمله‌ها و ریشه‌یابی کلمه‌ها به ابزارها و پردازش‌های دقیق‌تری نیازمند است.

پالایش‌های اولیه‌ای روی متن‌ها نیاز است انجام شود تا برای نمونه کلمه‌های با تکرار بسیار کم در تعداد زیادی متن حذف شوند، زیرا احتمال دارد این کلمات با فراوانی کم، خطای املایی باشند. همزمان باید دقت نمود گاهی برخی از کلمه‌های با فراوانی کم کلمه‌های درستی هستند که نقش تعیین کننده‌ای نیز در برخی از کاربردها مانند خوشه‌بندی متن دارند. زمینه‌های گوناگونی از علوم و مهندسی کامپیوتر و زبان‌شناسی همچون پردازش زبان‌های طبیعی، بازیابی اطلاعات و متن‌کاوی به بررسی استخراج و تحلیل باهم‌آیی‌هایی کلمه‌ها وابسته است [۲].

داسیلوا و لوپز (۱۹۹۹) رویکردی را برای گسترش معیارهای گوناگون با هم‌آیی پیشنهاد و پیاده سازی کردند که باهم‌آیی‌های nتایی[2] با n>2 را به دو بخش تقسیم کردند و آنها را به عنوان شبه باهم‌آیی‌های دوتایی[3] در نظر گرفتند [۳]. باهم‌آیی‌های دو کلمه‌ای را اغلب به کمک ضرایب تخمین وابستگی (AM[4]) مقایسه می‌کنند [۲]. ساده‌ترین روش گسترش AMها گسترش یک مقدار تنها در یک بُعد است. این روش توسط تادیچ و همکاران (۲۰۰۳) به کار گرفته شد [۴]. روش ارائه شده توسط مک‌اینس (۲۰۰۴) از روش‌های مختلف برای گسترش یک AM استفاده کرد، اما این روش‌ها تنها برای لگاریتم احتمال‌ها[5] استفاده شدند. بعد از محاسبه این مقدار برای هر مدل، مدلی که به بهترین شکل باهم‌آیی‌های nتایی را بیان کند به عنوان مقدار AM برای این باهم‌آیی‌ها انتخاب می‌شود [۵]. دین (۲۰۰۵) از نسبت مرتبه[6] برای نرمالیزه کردن بین باهم‌آیی‌های nتایی با طول‌های مختلف استفاده کرد و آنها را قابل مقایسه کرد [۶]. پترویچ و همکاران (۲۰۰۶) علاوه بر گسترش AMهای مختلف با روش مستقیم، یک روش ابتکاری AM برای باهم‌آیی‌های سه‌تایی[7] پیشنهاد دادند. این روش ابتکاری بر پایه مقدار اطلاعات مشترک است که اطلاعاتِ الگوی POS[8] مربوط به باهم‌آیی‌های nتایی را در نظر می‌گیرد [۷]. سرتان (۲۰۰۸) یک چارچوب روش‌شناسی برای

---

1 Collocation
2 N-Gram
3 Bigram
4 Association Measure
5 Log Likelihood
6 Rank Ratio
7 Trigram
8 Part Of Speech

شناسایی مبتنی بر نحو، باهم‌آیی‌های نامزد در متن منبع را قبل از مرحله محاسبات آماری فراهم کرد. او این روش استخراج را بر روی چهار زبان انگلیسی، فرانسوی، اسپانیایی و ایتالیایی ارزیابی کرد. این روش بر اساس اعمال محدودیت‌های نحوی بر روی اجزاء به جای محدودیت‌های مجاورت خطی است [۸]. کالسون (۲۰۱۰) و همکاران به جای استفاده از مقداردهی‌های آماری قبل، این کار را بر پایه الگوریتم‌های مجاورت انجام دادند [۹]. شیجون و همکاران (۲۰۱۵) به ساخت پایگاهی از باهم‌آیی‌های معنایی پرداختند و از قواعد معنایی برای این کار استفاده کردند [۱۰]. کائو و همکاران (۲۰۱۵) با یکپارچه سازی دانش نحوی و معنایی، یک ابزار استخراج سه لایه را پیشنهاد کردند. این کار در ابتدا به پیدا کردن باهم‌آیی‌های جانبی بر اساس تکرار در لایه اول می‌پردازد. در لایه دوم با استفاده از دانش نحوی به استخراج باهم‌آیی‌های نیمه جانبی[1] می‌پردازد. و در آخرین لایه با توجه به دانش معنایی آن‌ها را استخراج می‌کند [۱۱]. انگوین و همکاران میزان یادگیری دانشجویان ویتنامی رشته زبان انگلیسی را در باهم‌آیی‌های فعل-اسم، اسم-صفت بررسی و پیش‌بینی کردند و همچنین باهم‌آیی‌هایی را که تکرار بسیار بالایی از نظر معیارهای مختلف داشتند بررسی و پیش‌بینی کردند [۱۲]. وو و همکاران یک سیستم یادگیری باهم‌آیی‌ها را طراحی و توسعه دادند که از متن مقاله‌های ویکی‌پدیا ساخته شده است. این سیستم یک رابط کاربری مناسب را برای کاربران ارائه می‌دهد و باهم‌آیی‌ها را با توجه به الگوی نحوی و تکرار، مرتب می‌کند [۱۳]. شوجی و همکاران (۲۰۱۶) به ارائهٔ یک مرور کلی از ابزارهای تحلیل باهم‌آیی برای یادگیری زبان چینی پرداختند [۱۴]. اسپینوزا و همکاران (۲۰۱۶) به روابط باهم‌آیی‌ها در WordNet پرداختند که WordNet یکی از منابع واژگانی برای پردازش زبان طبیعی است. آن‌ها ColWordNet(CWN) را معرفی کردند که نسخهٔ گسترش‌یافته WordNet با اطلاعات دقیق از باهم‌آیی‌ها است [۱۵]. ورما و همکاران (۲۰۱۶) روشی را برای تجزیه و تحلیل باهم‌آیی‌ها ارائه دادند که محدودیت‌های روش‌های آماری را ندارد [۱۶]. گارسیا و همکاران (۲۰۱۷) یک روش جدید را برای استخراج باهم‌آیی‌های چند زبانه ارائه دادند که از مفاهیم موازی برای یادگیری لغات دو زبانه استفاده نمودند [۱۷].

پارک و همکاران (۲۰۱۶) به مشکلات تجربی استخراج باهم‌آیی‌ها در زبان ژاپنی و کره‌ای پرداختند و استانداردی را برای شناسایی باهم‌آیی‌ها در این زبان‌ها ارائه دادند. آن‌ها به بحث درباره تحلیل آماری الگوهای باهم‌آیی نپرداختند و بر جنبه‌های تجربی پژوهش تمرکز کردند [۱۸]. داس (۲۰۱۲) به استخراج باهم‌آیی‌های کلمات در مجموعه‌ای از متن‌های زبان بنگالی پرداخت [۱۹]. همچنین ابراهیم‌زاده و همکاران باهم‌آیی‌های دوتایی و سه‌تایی را از متن فارسی بایگانی روزنامه همشهری در یک بازه زمانی خاص استخراج کردند و از نظر آماری و تجربی مورد بحث قرار دادند [۲۰]. بنابراین باهم‌آیی‌های به‌دست آمده برای کلمات به مجموعهٔ متنی به کار گرفته شده، منبع خبری (یا هر پایگاه متنی دیگر)، زبان نوشتاری، نوع متن (کتاب، وب‌گاه، مقالات و یا صوت) و محتوای موضوعی مجموعهٔ بررسی شده بستگی دارد. همچنین زمان نوشته شدن متن‌های مجموعه بر روی نتایج باهم‌آیی‌ها اثر دارد زیرا در بازه‌های زمانی گوناگون نام‌های متفاوتی به صورت متداول باهم می‌آیند برای نمونه در موضوع خبرهای سیاسی نام‌های ترکیبی رییس‌جمهورها و وزیران آنها بیشتر باهم می‌آیند یا این که موضوع یا موضوع‌هایی تیتر خبرهای روزانه یا ماهانه می‌شود. پس هر کدام از پژوهش‌های انجام شده در زمینهٔ باهم‌آیی کلمات، جدای از به کارگیری الگوریتم‌های متفاوت به متن‌های به کار گرفته شده بر روی آنها نیز وابسته است و نتیجه‌های تا اندازه‌ای گوناگون و سودمند را به وجود می‌آورد. در این مقاله نیز مجموعهٔ متفاوتی از متن‌ها را برای استخراج باهم‌آیی‌ها به کار بردیم و باهم‌آیی‌های خبرهای انگلیسی وب‌گاه رادیوی صدا و سیمای جمهوری اسلامی ایران[2] را در بازهٔ زمانی ۱۳۸۶ تا ۱۳۸۹ به‌دست آوردیم [۲۱].

پترویچ و همکاران (۲۰۰۹) روش‌های به کار گرفته شده توسط تادیچ (۲۰۰۳) ، داسیلوا و لوپز (۱۹۹۹) و مک اینس (۲۰۰۴) را به قانونی کلی تبدیل کردند که برای هر AM قابل استفاده باشند [۲].

چون قالب‌های صفحه‌های HTML وب‌گاه انگلیسی صدا و سیمای جمهوری اسلامی به یک شکل نیستند، برای به‌دست آوردن متن خالص از صفحه‌های آنها برنامه‌ای نوشتیم تا کدهای HTML را بررسی کند و در صورت نیاز آنها را ویرایش کند و جمله‌ها و پاراگراف‌های مناسب را بیابد. همچنین در این پژوهش باهم‌آیی‌ها را روی متن ریشه‌یابی شده و ریشه‌یابی نشده انجام دادیم و نتایج آن‌ها را به صورت جداگانه به دست آورده و مقایسه کردیم.

---

[1] Semi-peripheral

[2] http://english.irib.ir/news

۲- تعریف‌ها و مفهوم‌های پایه در باهم آیی

باهم‌آیی تعریف‌های فراوانی دارد که در ادامه به دو نمونه تعریف آن اشاره می‌کنیم:

۱. عبارتی از دو یا بیشتر از دو کلمه برای رساندن یک مفهوم.

۲. (تعریف قدیمی‌تر) کلمه‌های پشت سر هم و مرسوم و همیشگی در یک زبان دارای یک بار معنایی یا نحوی [۲۲]. برای نمونه international organizations و Islamic Republic of Iran دو باهم‌آیی به‌دست آمده از این مقاله است.

باهم‌آیی در بازیابی متن‌های ناقص (پردازش صفحات کتاب‌ها یا مجلات)، ترجمهٔ متن (لغت‌نامه‌های تخصصی که برای ترکیبات کلمات، معانی متناسب و بر پایه گفتگوی زبانی مردم دارند، مثل وب‌گاه ترجمه گوگل)، متن‌کاوی، تولید زبان طبیعی برای ربات‌ها و ساختِ ربات‌های سخنگو (ربات‌هایی که به‌جای استفاده و ترکیب کلمات واحد، از مجموعه‌های باهم‌آیی‌ها جهت ارتقای سطح گفتگوی خود استفاده کنند) کاربرد دارد [۲۲].

باهم‌آیی کلمات در چندین حوزه مورد بررسی قرار می‌گیرد که از جمله می‌توان به سه حوزه کلی نحوی، معنایی و آماری اشاره کرد. باهم‌آیی به لحاظ صرفی نتیجهٔ فرآیند ترکیبی واژه‌سازی محسوب می‌شود که در طول آن نقش نحوی کلمه نیز مورد نظر قرار می‌گیرد [۲۳]. به طور کلی در حوزه نحوی جمله توجه می‌شود. برای نمونه جمله «میز را بچین یا آماده کن» را در زبان انگلیسی به صورت Set the table می‌گویند. در صورتی که یک فرد فارسی‌زبان بخواهد همین جمله را به فارسی ترجمه کند احتمال دارد آن را «میز را تنظیم کن» ترجمه کند در حالی که ترجمهٔ مناسبی نیست. بنابراین ترجمهٔ کلمه به کلمه دارای اشکالاتی هست. باهم‌آیی‌های به دست از تعداد زیادی متن برای هر زبان به بهبود ترجمه کمک می‌کند [۲۲].

حوزه معنایی باهم‌آیی‌ها به شناسایی معنایی گروهی از کلمه‌های کنار هم کمک می‌کند که معنای آن‌ها روی هم متفاوت از معنای تک تک آن کلمه‌ها است و یک اصطلاح را تشکیل می‌دهند. برای نمونه معنی تک تک کلمه‌ها در ترکیب red tape «نوار قرمز» است ولی با در نظر گرفتن باهم‌آیی معنایی، «مقررات دست و پاگیر» و یا «فرمالیته اداری» به دست می‌آید [۲۴].

حوزه آماری به احتمال وقوع هر کلمه در یک باهم‌آیی و یا احتمال وقوع یک باهم‌آیی در یک متن می‌پردازد. بر اساس احتمال وقوع و مقدار فراوانی‌های به‌دست آمده و سایر پارامترهای آماری، همچون انحراف معیار۵، برای هر باهم‌آیی می‌توان مقادیر و پارامترهایی به‌دست آورد که بتوان باهم‌آیی‌های هر متن را دسته‌بندی و رتبه‌بندی کرد [۲۵]. ساسا پتروویچ یکی از کسانی است که به کمک ضرایب تخمین وابستگی و تعمیم آنها و نیز روش‌های ابتکاری موجود برای این ضرایب، فرمول‌های جدید و قابل قبولی را برای باهم‌آیی‌های بیش از دو کلمه به‌دست آورد [۲].

در این پژوهش، حوزه‌های آماری و نحوی باهم‌آیی را در نظر گرفتیم و به نقش کلمات در بررسی باهم‌آیی‌های سه‌تایی نیز توجه کردیم. در این مقاله، نخست قاعده‌های استخراج باهم‌آیی‌ها را بررسی کردیم سپس به چگونگی محاسبه ضرایب تخمین وابستگی برای باهم‌آیی‌های دوتایی و تعمیم این ضرایب برای باهم‌آیی‌های سطح بالاتر پرداختیم.

۳- مراحل کار

زبان برنامه‌نویسی پایتون را برای اجرای مراحل استخراج باهم‌آیی‌ها به همراه بسته‌های NLTK[1] و numpy به کار بردیم. مراحل کار، شامل ۵ قسمت اصلی است که به ترتیب در زیر بیان شده است. پس از به دست آوردن فهرستی از باهم‌آیی‌ها و ضرایب می‌توان روی آنها تحلیل‌های زبان‌شناسی نیز انجام داد که ورای این پژوهش است.

۱. پیش‌پردازش متن

۲. تعیین اسامی و نقش کلمات در جمله (POS)

۳. ریشه‌یابی متن

[1] Natural Language Toolkit

۴. استخراج باهم‌آیی‌ها (دوتایی و سه‌تایی)
۵. تعیین ضرایب تخمین وابستگی برای باهم‌آیی‌های به‌دست آمده

۳-۱- پیش‌پردازش متن

حدود ۴۲۰۰۰ خبر در پایگاه داده ما وجود دارد که برخی از این متن‌ها، اطلاعات سودمندی را در بر ندارند و با حذف این متن‌ها، خبرهای یکپارچه‌ای را به دست آوردیم. خبرهای حذف شده عبارتند از:

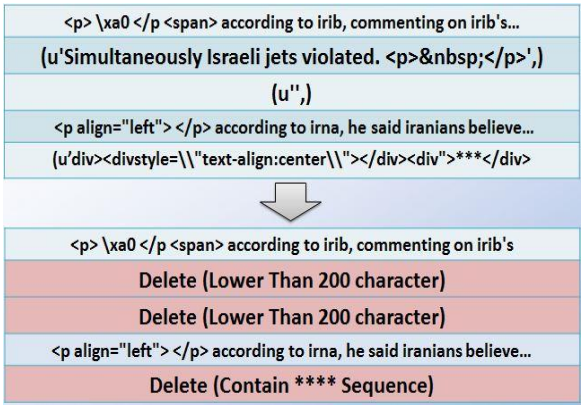
۱. متن‌های کمتر از ۲۰۰ حرف: احتمالا مربوط به دانلود یا عکس یا پرونده صوتی و تصویری یا همانند آن هستند که در پردازش متنی این پژوهش، ارزش چندانی ندارند.
۲. متن‌های ستاره‌دار: پس از بررسی تعدادی از خبرها دیدیم که خبرهای درون ستاره ارزش خبری ندارند.
۳. برچسب‌های HTML: نخست کتابخانهٔ HTMLParser را در پایتون به کار بردیم تا *برچسب‌های* ناقص را اصلاح کنیم تا در کار حذف آنها به مشکلی بر نخوریم. سپس این برچسب‌ها را حذف کردیم و باز خبرهای کمتر از ۱۰۰ حرف را نیز حذف کردیم.

تعداد خبرهای پایگاه داده پس از حذف این خبرها از ۴۲۰۰۰ به ۲۴۰۰۰ خبر کامل کاهش یافت. جدول ۱ چگونگی کاهش رکوردهای مجموعه داده‌ها پس از پیش‌پردازش را نشان می‌دهد.

جدول ۱: اثر پیش‌پردازش بر مجموعه داده‌ها

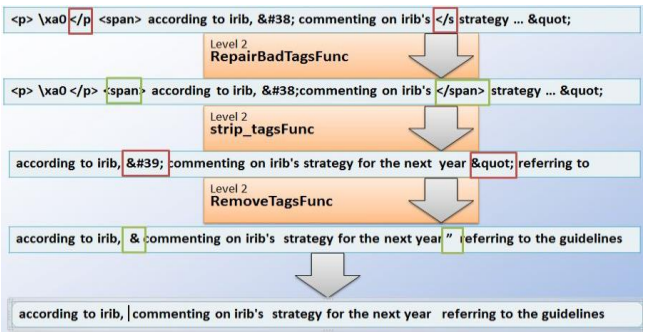
	تعداد خبر	تعداد کل کلمات
قبل از پیش‌پردازش	۴۲۰۰۰	۸ میلیون
بعد از پیش‌پردازش	۲۴۰۰۰	۵.۵ میلیون
درصد اطلاعات مناسب باقیمانده	٪۵۷	٪۶۸

طبق جدول ۱ حدود ۴۳ درصد خبرها از لحاظ محتوایی مشکل داشتند که در مرحله پیش‌پردازش آنها را حذف کردیم. بر اساس تعداد کلمات حذف شده در این خبرها نسبت به کل کلمات در همهٔ خبرها، حدود ۳۲ درصد کلمات ارزش متنی نداشتند. شکل ۱ مراحل اجرای این پیش‌پردازش را برای چند نمونه نشان می‌دهد.



شکل ۱: مراحل حذف اطلاعات نامناسب از مجموعه داده‌ها

در شکل ۱ از ۵ خبر نمونه، سه مورد حذف شد. کمبود تعداد حروف در دو متن حذف شدهٔ نخست باعث حذف آنها شد و داشتن دنباله‌ای از ستاره‌ها سومین را حذف کرد. شکل ۲ نیز چگونگی اصلاح و حذف برچسب‌های HTML را نشان می‌دهد.



شکل ۲: مرحله ترمیم و حذف برچسب‌ها در متن

۳-۲- شناسایی نام‌ها و نقش کلمات در جمله

کلمه‌ها و ترکیب‌هایی را ریشه‌یابی می‌کنیم که اسم خاص درون آنها نباشد. برای نمونه Islamic Revolution یک اسم است و نباید آن را به Islam revolve ریشه‌یابی نمود.

توابع برچسب‌گذار NLTK را به کار گرفتیم تا بر پایهٔ یک سری قواعد و کلمات پیش‌فرض در پایگاه دادهٔ NLTK نقش کلمات را تعیین کنیم. NLTK به هر کدام از این نقش‌ها یک حرف اختصاری را نسبت می‌دهد که بیانگر نوع کلمه است و برای پژوهش کنونی نقش‌های زیر کافی است:

- اسامی با برچسب‌های Noun = N
- صفت‌ها با برچسب‌های Adjective = A
- ایست‌واژه‌ها[1] با برچسب‌های Stopwords= S
- و بقیه برچسب‌ها Other = X

جدول ۲ نتایج به‌دست آمده از این مرحله را نشان می‌دهد. تعداد کلمه‌ها در پایگاه داده ۵۳۴۴۷۲۵ است.

جدول ۲: درصد نوع کلمات در متن

	اسم	صفت	ایست‌واژه	غیره
تعداد کلمات	۱۷۴۴۶۵	۳۶۲۳۰۴	۲۳۰۶۴۰۱	۹۳۱۹۵۵

گرچه اسم‌ها در برچسب‌گذار شناسایی شدند ولی نام‌های مرکب را نیز باید شناسایی کرد. برای نمونه: President Dr. Mahmoud Ahmadi Nejad

خروجی Tagger برای این ورودی به این شکل است: President=N, Dr.=N, Mahmoud=N, Ahmadi=N, Nejad=N

تابع Chunker سه کلمهٔ Mahmoud Ahmadi Nejad را به نام مرکب یک شخص تبدیل می‌کند. این تابع بر پایه POSها و معنی برخی کلمه‌ها، هویت کلمه‌هایی همچون شخص، مکان، ارگان و غیره را شناسایی می‌کند.

تابع برچسب‌گذار معایبی دارد، برای نمونه علامت‌های مانند پرانتز باز را گاهی به عنوان اسم در نظر می‌گیرد. برای حل این مشکل فهرستی از علامت‌های درون زبان را به‌دست آوردیم و در دستهٔ X گذاشتیم تا تابع برچسب گذار آنها را به عنوان اسم در نظر نگیرد.

مشکل دیگر زمان‌گیر بودن اجرای این تابع‌ها است. در این پژوهش، در آزمایش‌هایی که انجام دادیم اجرای این دو تابع این دو تابع حدود ۶ تا ۱۲ ساعت (بسته به پارامترهای گوناگون) به طول انجامید که در مقام مقایسه این زمان بسیار کمتر از محاسبهٔ باهم‌آیی‌ها (چندین روز) است.

[1] Stop Words

ایست‌واژه‌ها کلمه‌هایی هستند که در متن ها و گفتگوها زیاد به کار می‌روند و بر اساس قاعده‌های زبان شناسی انتخاب می‌شوند.برای نمونه در زبان انگلیسی ضمایر فاعلی و ربطی مثل ایست‌واژه‌ها کلمه‌هایی هستند که در متن ها و گفتگوها زیاد به کار می‌روند و بر اساس قاعده‌های زبان شناسی انتخاب می‌شوند.برای نمونه در زبان انگلیسی ضمایر فاعلی و ربطی مثل and، or و یا افعالی همچون am، is و are که زیاد در متن‌ها دیده می‌شوند ایست‌واژه هستند. فهرست واحدی از ایست‌واژه‌ها وجود ندارد زیرا الگوریتم‌های شناسایی آنها و نیز قاعده‌های گوناگون زبان‌شناسی که این کلمات را انتخاب می‌کنند روش یکسانی ندارند. ایست‌واژه‌ها را به دو دسته تقسیم کردیم:

۱. کلمات پیش‌فرضی که NLTK به ما می‌دهد.
۲. کلماتی که تابع برچسب‌گذار آنها را ایست‌واژه تشخیص دهد.

۳-۳- ریشه‌یابی متن

اصطلاح ریشه‌یابی به معنای حذف پسوندها، پیشوندها و میانوندها و به طور کلی قسمت‌های اضافی کلمه برای به‌دست آوردن ریشه کلمه است. از جمله اهداف این کار در بازیابی اطلاعات، جستجوی کلمه بر اساس ریشهٔ آن می‌باشد [۲۶]. این کار تا حدی باعث بهبود باهم‌آیی‌ها و متمرکز شدن نتایج آن می‌گردد. الگوریتم‌های ریشه‌یابی در دو دسته کلی الگوریتم‌های مبتنی بر لغت‌نامه و الگوریتم‌های مبتنی بر قانون تقسیم‌بندی می‌شوند که ما از الگوریتم مبتنی بر قانون Lancaster استفاده می‌کنیم که نسبت به انواع قدیمی‌تر آن مثل پورتر [۲۷] عملکرد بهتری دارد.

کامل نبودن ریشه‌یابی یکی از مشکلات است که از آن صرف‌نظر می‌کنیم زیرا دقت الگوریتم ریشه‌یابی مورد استفاده هرچند کامل نیست ولی تضمین خوبی برای ریشه‌یابی کلمات به ما می‌دهد. در زیر یک نمونه از ریشه‌یابی را مثال زده‌ایم:

یک بخش از جمله پیش از ریشه‌یابی:

AllText = **[**According**,X]** **[**to**,S]** **[**IRIB**,N]** **[**,**,X]** **[**commented**,X]** **[**powerful**,A]** **[**Strategy**,N]** **[**referring**,N]** **[**to**,S]** **[**Predident**,N]** **[**Dr. **,N]** **[**Mahmoud Ahamadi Nejad**,N]** **[**. **,X]**

جمله بالا بعد از ریشه‌یابی:

AllTextStem = **[**Accord**,X]** **[**to**,S]** **[**IRIB**,N]** **[**,**,X]** **[**comment**,X]** **[**pow**,A]** **[**Strategy**,N]** **[**referring**,N]** **[**to**,S]** **[**President**,N]** **[**Dr. **,N]** **[**Mahmoud Ahamadi Nejad**,N]** **[**. **,X]**

۳-۴- استخراج باهم آیی‌ها

بعد از انجام مرحله‌های پیشین، متن‌های به‌دست آمده را به هم متصل می‌کنیم و در یک رشته قرار می‌دهیم که آماده استخراج باهم‌آیی است. متن‌های به‌دست آمده از مراحل گفته شده در دو دسته تقسیم می‌شوند که شامل متن اصلی و متن ریشه‌یابی شده‌اند و باهم‌آیی‌ها نیز در دو گروه ریشه‌یابی شده و ریشه‌یابی نشده تقسیم می‌شوند. باهم‌آیی‌ها بر اساس تعداد کلمات تشکیل‌دهندهٔ آن به گروه‌های زیر تقسیم می‌شوند:

- دوتایی‌ها: ترکیبی از دو کلمه. برای نمونه:

different political
Dr. Mahmoud Ahmadi Nejad

نمونهٔ اول دربردارندهٔ دو کلمه، ولی نمونهٔ دوم بیش از دو کلمه است. این مورد به خاطر این است که Mahmoud Ahmadi Nejad یک کلمه است که در مرحلهٔ شناسایی اسم‌ها این عبارت را به عنوان یک کلمه و یک اسم در نظر گرفته‌ایم. پس ممکن است باهم‌آیی دوتایی شامل بیش از دو کلمه نیز باشد.

- سه‌تایی‌ها: ترکیبی از سه کلمه. برای نمونه:

peaceful nuclear activities
President Dr. Mahmoud Ahmadi Nejad
Supreme Leader of Islamic Revolution

- باهم‌آیی‌هایی با تعداد کلمه‌های بیشتر، که در کل آنها را Ngram می‌شناسند. این نوع باهم‌آیی‌ها شامل زنجیره‌ای از N کلمه می‌باشند [۲].

در مقاله حاضر ما فقط باهم‌آیی‌های دوتایی و سه‌تایی را استخراج می‌کنیم؛ در بیشتر مقاله‌ها نیز به همین دو بسنده می‌کنند. برای استخراج باهم‌آیی‌ها روش‌های زیادی وجود دارد که بر پایهٔ شرایط و اهداف پروژه می‌توان تعدادی از آنها را برگزید. قانون‌های زیر را برای استخراج باهم‌آیی‌ها در این مقاله به کار بردیم:

۱. استخراج نکردن باهم‌آیی‌هایی که تعدادِ تکرارِ کم دارند: در این مورد هیچ اتفاق‌نظری میان پژوهشگران نیست که این عدد (آستانه) چقدر باید باشد و چه مقدار از باهم‌آیی‌ها را پوشش دهد؟ در اجرای اولیه آستانه‌ای را در نظر نگرفتیم زیرا هدفِ اولیه استخراج تمامی باهم‌آیی‌ها بود. در بیشتر متن‌ها، تعداد باهم‌آیی‌های با تکرار ۱ یا ۲ یا ۳، حدود ۸۰ درصد کل باهم‌آیی‌ها را تشکیل می‌دهند. به همین خاطر زمانی که این آستانه انتخاب شود ۸۰ درصد باهم‌آیی‌ها رد می‌شوند و سرعت اجرای برنامه بالا می‌رود. در بیشتر کارهای پژوهشی در این زمینه، آستانه‌ای را برای استخراج باهم‌آیی‌ها قرار می‌دهند تا هم سرعت اجرای برنامه بالا رود و هم اینکه باهم‌آیی‌های با تکرار کم که ارزش پردازشی ندارند در نتیجه‌ها نیایند.

۲. بر پایهٔ مقاله ساسا پتروویچ، اجزای باهم‌آیی‌ها فقط باید دربردارندهٔ یکی از نقش‌های اسم، صفت و ایست‌واژه باشند یا به عبارت دیگر، باهم‌آیی نباید دربردارندهٔ کلمه‌ای با POS(X) باشد [۲].

۳. باز هم بر پایهٔ همان مقاله، ایست‌واژه‌ها نباید در اول و آخر باهم‌آیی‌ها باشند ولی در هر جای دیگر باهم‌آیی، می‌توانند باشند[۲]. پس باهم‌آیی‌های دوتایی هیچ‌گاه دربردارندهٔ ایست‌واژه نیستند ولی در باهم‌آیی‌های سه‌تایی به شکل زیر می‌توانند وجود داشته باشند:

ASA, ASN, NSN, NSA

در گام بعدی باید پارامترهای مورد نیاز را برای محاسبهٔ هر کدام از ضرایب تخمین وابستگی به دست آوریم. جدول ۳ و جدول ۴ چگونگی محاسبهٔ این پارامترها را نشان می‌دهند [۲۸].

جدول ۳: پارامترهای ضرایب تخمین وابستگی برای دوتایی‌های UV

	V=v	V≠v	
U=u	O_{11}	O_{12}	$O_{11}+O_{12}=R_1$
U≠u	O_{21}	O_{22}	$O_{21}+O_{22}=R_2$
	$C_1 = O_{11}+O_{21}$	$C_2 = O_{12}+O_{22}$	$R_1+R_2= C_1+C_2=N$

جدول ۴: محاسبه پارامترهای ضرایب تخمین وابستگی برای دوتایی‌های UV

	V=v	V≠v
U=u	$E_{11} = (R_1C_1)/N$	$E_{12} = (R_1C_2)/N$
U≠u	$E_{21} = (R_2C_1)/N$	$E_{22} = (R_2C_2)/N$

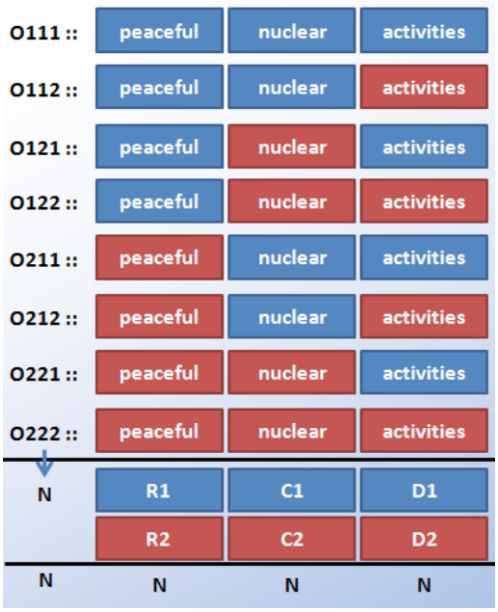
برای هر باهم‌آیی دوتایی UV (اول U و بعد V)، ۴ پارامتر O_{ij} را می‌توان به‌دست آورد که:

- O_{11} به این معنی است که کلمه اول باهم‌آیی دقیقاً u و کلمه دوم باهم‌آیی دقیقاً v باشد.
- O_{12} یعنی کلمه اول دقیقاً u و کلمه دوم باهم‌آیی هر چیزی غیر از v باشد.
- O_{21} یعنی کلمه اول باهم‌آیی هر چیزی غیر از u و کلمه دوم دقیقاً v باشد.
- O_{22} یعنی کلمه اول باهم‌آیی هر چیزی غیر از u و کلمه دوم هر چیزی غیر از v باشد. حال اگر تمام این ۴ پارامتر را با هم جمع کنیم تعداد کل کلمات متن به‌دست می‌آید که آن را با N نشان می‌دهیم.

پارامتر R1 به معنای تعداد تکرار کلمه اول باهم‌آیی (u) در متن و R_2 به معنای تعداد کلمات متن به جز کلمه اول باهم‌آیی (u) است. C_1 و C_2 مانند R_1 و R_2 هستند با این تفاوت که برای کلمهٔ دوم باهم‌آیی (v) است. پارامترهای E_{ij} نیز پارامترهایی برای محاسبه برخی از ضرایب تخمین وابستگی هستند که در بخش‌های بعدی توضیح می‌دهیم.

ماهیت پارامترها در باهم‌آیی‌های سه‌تایی همچنان همان است، ولی به دلیل وجود سه کلمه تا حدی نحوه به‌دست آوردن آنها متفاوت خواهد بود. مقدار یک داشتن هرکدام از i و j و k ها در پارامتر O_{ijk} به معنای حضور آن کلمه در باهم‌آیی است و مقدار دو داشتن آنها به معنای عدم حضور آن کلمه در باهم‌آیی است. شکل ۳ نحوه محاسبهٔ پارامترها را روی یک مثال نشان می‌دهد.

سه‌تایی‌ها نیز همچون دوتایی‌ها که پارامترهای R و C داشتند، پارامتر دیگری به نام D دارند که مربوط به تعداد تکرار کلمه سوم باهم‌آیی است.



شکل ۳: نحوه محاسبه Oijkها در باهم آیی‌های سه‌تایی

در زیر یک نمونه از استخراج باهم‌آیی‌ها را نشان می‌دهیم:

AllText = [According,N] [to,S] [IRIB,N] [,,X] [commented,X] [powerful,A] [Strategy,N] [referring,N] [to,S] [Predident,N] [Dr. ,N] [Mahmoud Ahamadi Nejad,N] [. ,X]

- سه تایی‌ها در متن ریشه‌یابی شده:

referring to President
Pow Strategy referring
President Dr. Mahmoud Ahmadi Nejad

- سه تایی‌ها در متن ریشه‌یابی نشده:

referring to President
Powerful Strategy referring
President Dr. Mahmoud Ahmadi Nejad

- دوتایی‌ها در متن ریشه‌یابی شده:

Pow Strategy
President Dr.
Dr. Mahmoud Ahmadi Nejad

- دوتایی‌ها در متن ریشه‌یابی نشده:

Powerful Strategy
President Dr.
Dr. Mahmoud Ahmadi Nejad

جدول ۵ تعداد باهم‌آیی‌های دوتایی و سه‌تایی از متن‌های ریشه‌یابی شده و ریشه‌یابی نشده را نشان می‌دهد.

جدول ۵: تعداد باهم آیی‌های استخراج شده

	دوتایی	سه‌تایی
متن ریشه‌یابی‌نشده	۲۹۹۸۲۹	۲۶۴۰۵۰
متن ریشه‌یابی شده	۳۲۷۴۷۹	۸۷۰۸۳

بعد از استخراج هر باهم‌آیی، تمامی اطلاعات مربوط به آن از جمله پارامترهای مربوطه و کلمات تشکیل‌دهندهٔ آن را در جداولی در پایگاه داده ذخیره می‌کنیم. برای سه‌تایی‌ها به ذخیره ۱۲ پارامتر نیاز داریم که ۸ تای آن مربوط به O_{ijk}ها و ۳ مورد هم برای نمایش کلمات تشکیل‌دهنده باهم‌آیی و یک پارامتر هم معرف این است که آیا کلمه وسط باهم‌آیی ایستواژه است یا خیر. کلمات ایستواژه می‌توانند در وسط باهم‌آیی حضور داشته باشند که با داشتن این پارامتر می‌فهمیم باهم‌آیی شامل ایستواژه است و در نتیجه در محاسبهٔ ضرایب تخمین وابستگی تغییراتی می‌دهیم.

۳-۵- تعیین ضرایب تخمین وابستگی

این ضرایب یک سری از فرمول‌های ریاضی هستند که با کمک اطلاعاتی که برای هر باهم‌آیی در اختیار داریم، مقداری عددی به آنها نسبت می‌دهیم تا بتوانیم باهم‌آیی‌ها را بر اساس این ضرایب مقایسه کنیم. از حدود ۸۵ فرمولی که در این زمینه دیده بودیم ۱۱ مورد پرکاربردتر را استفاده کردیم.

دایس مهم‌ترین ضریب AM می‌باشد که در سال ۱۹۹۳توسط *اسماجا* برای استخراج باهم‌آیی‌ها از بدنهٔ متن ها ابداع شده است که در سال ۱۹۹۹ شخصی به نام دایس این فرمول را برای باهم‌آیی‌های N-gram تعمیم داد. معادله ضریب دایس برای باهم‌آیی‌های دوتایی و سه‌تایی به شکل زیر می‌باشد:

$$Dice(Bigram) = \frac{2O_{11}}{R_1 + C_1} \qquad Dice(Bigram) = \frac{2O_{11}}{R_1 + C_1} \qquad (۱)$$

MI[1]: این ضریب بر اساس Maximum-LikeLihood است و هرچند تمایلش به باهم‌آیی‌های با مقدار کم است ولی در متون، مخصوصاً انگلیسی زیاد به کار برده می‌شود.

$$MI(BiGram) = Log\frac{O_{11}}{E_{11}} \qquad MI(BiGram) = Log\frac{O_{11}}{E_{11}} \qquad (۲)$$

تا به اینجا دو دسته فرمول را توضیح دادیم در ادامه به دیگر فرمول‌ها فقط اشاره می‌کنیم [۲۹,۳۰].

:Jacard

$$Jaccard(Trigram) = \frac{O_{111}}{O_{111} + O_{112} + O_{121} + O_{122} + O_{211} + O_{212} + O_{221}} \qquad Jaccard(Bigram) = \frac{O_{11}}{O_{11} + O_{12} + O_{21}} \qquad (۳)$$

:Gmean

$$Gmean(Trigram) = \frac{O_{111}}{\sqrt{R_1C_1D_1}} \qquad Gmean(Bigram) = \frac{O_{11}}{\sqrt{R_1C_1}} \qquad (۴)$$

:LiddleLL

$$LiddleLL(Bigram) = \frac{N(O_{11} - E_{11})}{C_1C_2} \qquad LiddleLL(Bigram) = \frac{N(O_{11} - E_{11})}{C_1C_2} \qquad (۵)$$

:Ods-Ratio-Disc

$$ODL(Trigram) = Log\frac{(O_{111} + 1/2)(O_{222} + 1/2)}{(O_{112} + 1/2)(O_{121} + 1/2)(O_{122} + 1/2)} \qquad ODL(Bigram) = Log\frac{(O_{11} + 1/2)(O_{22} + 1/2)}{(O_{12} + 1/2)(O_{21} + 1/2)} \qquad (۶)$$
$$(O_{211} + 1/2)(O_{212} + 1/2)(O_{221} + 1/2)$$

[1] Mutual Information

Chi-Squared:

$$CS(Bigram) = \frac{N(O_{11} - E_{11})^2}{E_{11}E_{22}} \qquad CS(Trigram) = \frac{N(O_{111} - E_{111})^2}{E_{111}E_{222}} \qquad (۷)$$

Chi-SquaredH:

$$CSH(Bigram) = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{R_1R_2C_1C_2} \qquad CSH(Bigram) = \frac{N(O_{111}O_{222} - O_{112}O_{121}O_{122}O_{211}O_{212}O_{221})^2}{R_1R_2C_1C_2D_1D_2} \qquad (۸)$$

Z-Score:

$$Z - Score(Bigram) = \frac{O_{11} - E_{11}}{\sqrt{E_{11}}} \qquad Z - Score(Trigram) = \frac{O_{111} - E_{111}}{\sqrt{E_{111}}} \qquad (۹)$$

T-Score:

$$T - Score(Bigram) = \frac{O_{11} - E_{11}}{\sqrt{O_{11}}} \qquad T - Score(Trigram) = \frac{O_{111} - E_{111}}{\sqrt{O_{111}}} \qquad (۱۰)$$

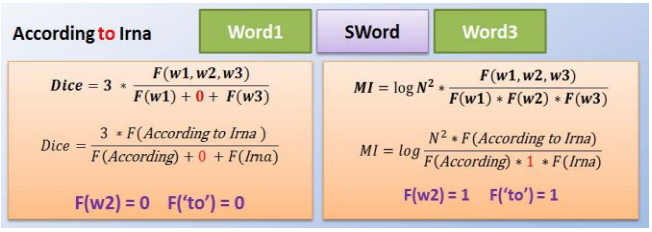
Local-MI:

$$Local - MI(Bigram) = O_{11}.Log\frac{O_{11}}{E_{11}} \qquad Local - MI(Trigram) = O_{111}.Log\frac{O_{111}}{E_{111}} \qquad (۱۱)$$

۳-۶- استثناء در محاسبه ضرایب تخمین وابستگی

عددهای کوچکی برای باهم‌آیی‌های دربردارندهٔ ایست واژه به دست می‌آید بنابراین این باهم‌آیی‌ها رتبهٔ پایین‌تری را در نتایج باهم‌آیی‌ها به دست می‌آورند. ساسا پتروویچ این مقدارِ کم را به این دلیل می‌داند که ایست‌واژه‌ها فراوانی بالایی در متن دارند و این تکرارِ زیاد مستقیما بر روی پارامتر O_{212} اثر می‌گذارد. در O_{212} کلمهٔ یکم و سوم باهم‌آیی هر چیزی غیر از u و z است ولی کلمه وسط باهم‌آیی دقیقاً v است. مقدار این عدد به‌طور تقریبی با تعداد تکرار کلمه v در متن (کلمه v یک کلمه ایست‌واژه و پرتکرار است) برابر است. در اغلبِ فرمول‌های ضریب‌های تخمین، این پارامتر در مخرج کسرها می‌آید بنابراین ضریب‌های این باهم‌آیی‌ها (دربردارندهٔ ایست واژه) مقدار بسیار ناچیزی می‌شوند. ساسا پتروویچ به جای پارامتر C_1 (تعداد تکرار کلمه دوم که در اینجا، این کلمه ایست‌واژه و پرتکرار است) مقدار صفر یا یک را قرار می‌دهد. به این شکل که اگر C_1 در ضرب ظاهر شده بود (به شرط اینکه کلمه دوم باهم‌آیی‌های سه‌تایی ایست‌واژه باشد) مقدار ۱ و اگر در جمع ظاهر شود به جای مقدار اصلی، مقدار صفر را جایگزین می‌کند.

با این کار ضریب محاسبه شده برای این نوع باهم‌آیی‌ها مقدارِ قابل قبولی را می‌گیرد و رتبهٔ خوبی را نیز به دست می‌آورد[۲]. برخی از ضرایب، پارامتر C_1 را ندارند بلکه پارامتر O_{212} را دارند که بر پایهٔ شیوهٔ یاد شده مانند پارامتر C_1 عمل می‌کنیم. هرچند با این روش نتایج خوبی به دست می‌آید ولی باهم‌آیی‌های دارای ایست واژه رتبهٔ بالاتری را نسبت به بعضی دیگر از باهم‌آیی‌های بدون ایست واژه پیدا می‌کنند. ما نیز مانند برخی پژوهش‌های دیگر در این زمینه از حل این مشکل صرف‌نظر کردیم. شکل ۴ نمونه‌ای از این باهم‌آیی را نشان می‌دهد:



شکل ۴: حل مشکل باهم آیی‌های شامل ایست واژه

در باهم‌آیی According to Irna در شکل ۴ کلمهٔ وسط باهم‌آیی (to) یک ایست‌واژه است و باید از قاعده‌ای که گفته شد ضرایب را محاسبه کنیم. جدول ۶ پارامترهای مربوط به این باهم‌آیی را نشان می‌دهد. پارامتر O_{212} برای این باهم‌آیی مقداری بسیار بیشتر از دیگر پارامترها دارد که باعث می‌شود مقادیر ضرایب تخمین وابستگی برای چنین باهم‌آیی‌هایی، بسیار ناچیز بشود. به همین منظور مقدار O_{212} را صفر یا یک درنظر می‌گیریم.

جدول ۶: پارامترهای باهم آیی‌های According to Irna

پارامترها	مقدار فراوانی
O_{111}	۹۴
O_{112}	۱۴۱۹
O_{121}	۱
O_{122}	۵
O_{211}	۲۶۶۴
O_{212}	۱۴۶۰۲۷
O_{221}	۳۷۸

۴- نتایج عملی

در نتیجه‌های به‌دست آمده مشکلاتی وجود دارد که برای هرکدام راه‌حل‌هایی پیشنهاد می‌کنیم.

نخستین مشکل، ساختار فرمول‌های تخمین وابستگی است. برای نمونه در فرمول دایس ۱۲ می‌توان تشخیص داد، مقداری که این ضریب برمی‌گرداند بزرگ‌تر از صفر و حداکثر یک است. $F(x)$ نشان‌دهنده فراوانی تکرار کلمه اول باهم‌آیی و $F(y)$ فراوانی تکرار کلمه دوم و نیز $F(xy)$ فراوانی تکرار باهم‌آیی تشکیل‌شده از دو کلمهٔ x و y است.

$$\text{Dice} = \frac{F(x) + F(y)}{2F(xy)} \qquad (۱۲)$$

اکنون اگر شرایط زیر را برای یک باهم‌آیی با تکرار یک در نظر بگیریم (یعنی باهم‌آیی که فقط یک‌بار در کل متن وجود داشته باشد) مقدار O_{11} برابر ۱ می‌شود و اگر کلمه‌های اول و دوم نیز در متن یک‌بار آمده باشند (یعنی کلمه اول و دوم باهم‌آیی، فقط یک‌بار در متن باشند و آن‌هم فقط در همان باهم‌آیی) آن‌گاه مقادیر O_{12} و O_{21} نیز برابر یک می‌شوند. بنابراین مقداری که دایس برای این باهم‌آیی می‌دهد برابر با یک است.

$$\text{Dice} = \frac{1+1}{2 * (1)}$$

این مقدار یعنی بیشترین مقدار ضریب برای یک باهم‌آیی و اگر بخواهیم باهم‌آیی‌های استخراج‌شده را بر اساس ضرایب دایس مرتب کنیم چنین باهم‌آیی‌هایی در صدر مقادیر قرار می‌گیرند. همچنین بسیاری از باهم‌آیی‌های موجود در متن به این شکل هستند؛ یعنی تعداد تکرار کلمه اول و دوم و تکرار باهم‌آیی شامل این دو کلمه، تقریباً برابر هستند. این نوع باهم‌آیی‌ها، ضرایب دایس با بیشترین مقدار را نتیجه می‌دهند که باعث خراب شدن قسمت تحلیل و بررسی می‌شوند. البته نه فقط برای دایس، بلکه برای بیشتر ضرایب تخمین وابستگی این نکته صدق می‌کند که برای حل این مشکل، آستانه را به کار می‌بریم.

پس از چند بار اجرای اولیهٔ برنامه و به کمکِ نتیجه‌های باهم‌آیی‌های به دست آمده، مقدار ۲۵ را برای آستانهٔ ضریب دایس مناسب یافتیم زیرا باهم‌آیی‌هایی که فراوانی زیر ۲۵ دارند، ضرایب دایس برابر یک را می‌دهند. این قاعده به صورت کلی نیست.

این مقدار آستانه برای تعدادی از ضرایب دیگر نیز مناسب است زیرا بیشتر ضرایب تخمین وابستگی به هم مرتبط هستند. این مقدار برای ضریب MI می‌تواند کمتر باشد ولی بهتر است این مقدار را بر اساس ضریب دایس بگیریم تا نتیجه بهتری داشته باشیم. نکته دیگر اینکه در اکثر مقالات این مقدار آستانه عددی حدود ۳ تا ۱۰ و در موارد خاص بیشتر از این مقدار هم گرفته می‌شود.

در نتایج ما، باهم‌آیی‌های با تکرار یک برابر با ۲۰۳۱ می‌باشند. اگر قانون آستانه را بر روی باهم‌آیی‌ها اعمال کنیم از ۳۰۰۰۰۰ باهم‌آیی به‌دست آمده فقط ۲۸۰۰ باهم‌آیی مورد قبول واقع می‌شوند و بقیه باهم‌آیی‌ها به عنوان باهم‌آیی‌ها نامناسب حذف می‌شوند. برای اینکه بدانیم چه حدی برای انتخاب باهم‌آیی‌ها مناسب است به اینگونه عمل کردیم:

تمامی باهم‌آیی‌هایی که مقدار ضریب دایس آنها برابر یک است را انتخاب و سپس باهم‌آیی که بیشترین مقدار ضریب O_{11} را دارد پیدا کرده و این مقدار را، مقدار آستانهٔ استخراج باهم‌آیی‌ها قرار می‌دهیم. در پایگاه دادهٔ ما از بین ۲۰۳۱ باهم‌آیی با مقدار ضریب دایس برابر ۱، بیشترین مقدار O_{11} برابر با ۲۵ است که ما نیز مقدار ۲۵ را آستانه قرار می‌دهیم و بعد از اعمال این مقدار، دیگر در نتایج، عدد دایس با مقدار یک را نمی‌بینیم.

ضریب MI کاملاً حساس به مقدار آستانه است و هر مقداری را که اعمال کنیم، باز بهترین ضرایب MI مربوط به پایین‌ترین مقدار تکرار است. معمولاً مقدار مناسب آستانه برای این ضریب را به صورت تجربی در نظر می‌گیرند و یا همان آستانه‌ای را که برای ضریب دایس در نظر گرفته شده است برای این ضریب اعمال می‌کنند[۷]. همان‌گونه که برای باهم‌آیی‌های دوتایی قانون آستانه و موارد مربوط به آن توضیح داده شد، این موارد برای سه‌تایی‌ها نیز تا حد زیادی صدق می‌کند.

جدول ۷ و جدول ۸ نمونه‌های باهم‌آیی‌ها و ضرایب دایس را پیش و پس از اعمال آستانه نشان می‌دهد. باهم‌آیی‌های با فراوانی یک ضریب دایس یک نیز دارند که با در نظر گرفتن آستانه از آنها چشم پوشی کردیم.

جدول ۷: قسمتی از ضرایب دایس پیش از اعمال آستانه

		O_{11}	O_{12}	O_{21}	Dice
Guideline.Colin	Toogood	۱	۰	۰	۱
Effart	Settlement.Ibra…	۱	۰	۰	۱
High-power	Ku-band	۱	۰	۰	۱

جدول ۸: قسمتی از ضرایب دایس پس از اعمال آستانه

		O_{11}	O_{12}	O_{21}	Dice
Zain	al-Abedin	۱۱۷	۱۰	۰	۰,۹۵۹۰۱
Strategic Arms	Redcution	۳۴	۰	۳	۰,۹۵۷۷۴
Deepwater	Horizon	۶۵	۲	۷	۰,۹۳۵۲۵

جدول ۹ و جدول ۱۰ ضرایب MI را پیش و پس از در نظر گرفتن آستانه نشان می‌دهد.

جدول ۹: قسمتی از ضرایب MI پیش از اعمال آستانه

		O_{11}	O_{12}	O_{21}	MI
Guideline.Colin	Toogood	۱	۰	۰	۲۲,۳۴۹۶۸
Effart	Settlement.Ibra	۱	۰	۰	۲۲,۳۴۹۶۸
High-power	Ku-band	۱	۰	۰	۲۲,۳۴۹۶۸

جدول ۱۰: قسمتی از ضرایب MI پس از اعمال آستانه

		O_{11}	O_{12}	O_{21}	MI
Mavi	Marmara	۱۱	۰	۰	۱۸,۸۹۰۲۵
Magnum	Opus	۱۱	۲	۰	۱۸,۶۴۹۲۴

۱۸,۵۳۳۷۶	۱	۱۰	۱۲	Hersh	Seymour

در نتایج استخراج باهم‌آیی‌ها بدون در نظر گرفتن ضرایب و فقط بر اساس فراوانی‌ها نتایجی به‌دست آمد که به عنوان نمونه، جدول ۱۱ و جدول ۱۲ سه تا از بیشترین فراوانی‌های هر دسته از باهم‌آیی‌ها را نشان می‌دهد.

جدول ۱۱: نمونه‌هایی از باهم آیی‌های دوتایی با فراوانی بیشتر

	دوتایی ریشه‌یابی نشده		دوتایی ریشه‌یابی شده	
۱	Islamic Republic	۳۹۱۸	Islamic Republic	۳۹۱۸
۲	Zionist regime	۳۱۷۹	Zionist regime	۳۱۷۵
۳	Press TV	۲۳۴۵	Press TV	۲۳۴۵

جدول ۱۲: نمونه‌هایی از باهم آیی‌های سه‌تایی با فراوانی بیشتر

	سه‌تایی ریشه‌یابی نشده		سه‌تایی ریشه‌یابی شده	
۱	Republic of Iran	۲۲۰۰	Republic of Iran	۲۱۹۹
۲	Islamic Republic of Iran	۲۱۶۴	Islamic Republic of Iran	۲۱۶۳
۳	UN Security Council	۶۴۵	UN Security Council	۶۴۵

۵- نتیجه‌گیری

در این مقاله، تعدادی از روش‌های شناسایی باهم‌آیی‌های کلمه‌ها را بررسی کردیم و از میان آنها تعدادی را برای شناسایی باهم‌آیی‌های دو تایی و سه تایی برگزیدیم. به موضوع‌هایی همچون قاعده‌های شناسایی ایست‌واژه و آستانه و مسائل مرتبط به آنها نیز پرداختیم. به کمکِ نرم‌افزارها و بسته‌های مناسبی در حوزه پردازش زبان‌های طبیعی روش‌های مورد نظر را پیاده سازی کردیم و بر روی متن اخبار انگلیسی وب‌گاه رادیوی صدا و سیما در بازهٔ زمانی ۱۳۸۶ تا ۱۳۸۹ اجرا کردیم و نتایجی را به دست آوردیم. سپس به کمک نتیجه‌های به دست آمده بهبودهایی را در برنامه انجام دادیم که از آن جمله در نظر گرفتن آستانه مناسب بود. سپس دوباره برنامه را بر روی خبرها اجرا کردیم و نتیجه‌های بهتری را برای هر حالت های دو تایی و سه تایی و با ریشه‌یابی و بدون ریشه‌یابی به دست آوردیم و باز تغییراتی را ایجاد کردیم و این چرخه را چند بار انجام دادیم تا نتایج به نسبت خوبی را از پایگاه داده به‌دست آوردیم.

نتایج باهم‌آیی در پردازش زبان‌های طبیعی، پردازش گفتار، موتورهای جستجو، خطایاب و ساخت فرهنگ لغت به کار می‌رود. با بررسی باهم‌آیی‌ها، تا اندازه‌ای می‌توان به قالب فکری و یا خبری یک وب‌گاه، یا کتاب یا همانند آن نیز پی برد. بازهٔ زمانی متن‌ها نیز بر روی نتیجه‌های باهم‌آیی اثر قابل توجهی دارد و به نوعی نشان دهندهٔ مهم‌ترین نکاتی است که در آن بازهٔ زمانی در آن منبعِ خاص بیشتر روی آن تکیه شده است.

ریشه‌یابی یا عدم ریشه‌یابی در این کار بر روی نتایج کار اثر چندانی نداشت زیرا بیشترِ باهم‌آیی‌های با فراوانی زیاد اسم‌های خاص بودند که ریشه‌یابی نمی‌شدند. در این مقاله، ریشه‌یابی باعث شد که باهم‌آیی‌ها متمرکزتر شوند یا به عبارتی از پراکندگی مشتقات یک کلمه جلوگیری شود گرچه باز هم اسم‌های خاص(بدون ریشه‌یابی) رتبهٔ بالاتری را در نتایج داشتند. چون نتیجه‌های قابل قبولی از این پژوهش به دست آمد از مشکلاتِ بخش ریشه‌یابی نسخهٔ به کار گرفته شده در ابزار NLTK چشم پوشی کردیم. برای ادامهٔ این پژوهش کارهای زیر را پیشنهاد می‌کنیم:

- بهبود ریشه‌یابی
- یافتن باهم‌آیی‌های چندتایی با ریشه‌یابی و بدون ریشه‌یابی
- استخراج باهم‌آیی‌ها از مجموعه‌های دیگر و مقایسه نتایج به‌دست آمده از آنها

- انجام تحلیل‌های زبان‌شناسی یا تحلیل‌های مرتبط بر روی نتایج این پژوهش

مراجع

[۱] پالمر، فرانک، *نگاهی تازه به معنی شناسی*؛ ترجمهٔ صفوی، کوروش، چاپ دوم، تهران، نشر مرکز، ۱۳۷۴.

[2] Petrovic, S. Snajder, J. "Extending Lexical Association Measures for Collocation Extraction", Journal of Computer Speech and Language, 2009.

[3] DaSilva, J, F. Lopes, G, P. "A Local Maxima Method and A Fair Dispersion Normalization for Extracting Multi-Word Units from Corpora", 6th Meeting on the Mathematics of Language, Orlando, pp. 369–381, 1999.

[4] Tadic, M. Sojat, K. "Finding Multiword Term Candidates in Croatian", Proceedings IESL2003 Workshop, pp. 102-107, 2003.

[5] McIness, B, T. "Extending The Log Likelihood Measure to Improve Collocation Identification", Master's thesis, University of Minnesota, 2004.

[6] Deane, P. "A Nonparametric Method for Extraction of Candidate Phrasal Terms", Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, pp. 605-613, 2005.

[7] Petrovic, S. Snajder, J. Bojana. Kolar. "Comparison of Collocation Extraction Measures for Document Indexing", 28th International Conference on Information Technology Interfaces, Publisher: IEEE, 2006.

[8] Sertan, V. "Collocation Extraction Based on Syntactic Parsing", Ph.D, thesis, University of Geneva, Geneva, Switzerland, 2008.

[9] Colson J-P. "Automatic Extraction of Collocations: A New Web-Based Method", Proceedings of JADT 2010 – Statistical Analysis of Textual Data, Roma, pp. 397-408, 2010.

[10] Shijun, L. Yanqiu, S. Lijuan, Z. Yu, D. "Construction of Semantic Collocation Bank Based on Semantic Dependency Parsing", 29th Pacific Asia Conference on Language (PACLIC 29), Shanghai, 2015.

[11] Cao, J. Li, D. Huang, D. "A Three-Layered Collocation Extraction Tool and Its Application in China English Studies", Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, Lecture Notes in Computer Science, Vol 9427, pp. 38-49, 2015.

[12] Nguyen, T, M, H. Webb, S. "Examining Second Language Receptive Knowledge of Collocation and Factors That Affect Learning", Language Teaching Research, Vol 21, Issue 3, pp. 298-320, 2016.

[13] Wu, S. Li, L. Witten, I. Yu, "A. Constructing A Collocation Learning System from The Wikipedia Corpus", International Journal of Computer-Assisted Language Learning and Teaching, Vol 6, Issue 3, 2016.

[14] Shouji, L. Shulun, G. "Collocation Analysis Tools for Chinese Collocation Studies", Journal of Technology and Chinese Language Teaching, Vol 7, pp. 56-77, 2016.

[15] Espinosa-Anke, L. Camacho-Collados, J. Rodríguez-Fernández, S. Saggion, H. Wanner, L. "Extending WordNet with Fine-Grained Collocational Information Via Supervised Distributional Learning", Proceedings of COLING 2016: Technical Papers, The 26th International Conference on Computational Linguistics, Osaka, Japan, pp. 3422-3432, 2016.

[16] Verma, R.Vuppuluri, V. Nguyen, A. Mukherjee, A. Mammar, G. Baki, S. Armstrong, R. "Mining The Web for Collocations: IR Models of Term Associations", Proceedings of 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING), 2016.

[17] Garcia, M. Garcia-Salido, M. Alonso-Ramos, M. "Using Bilingual Word-Embeddings for Multilingual Collocation Extraction", Proceedings of the 13th Workshop on Multiword Expressions (MWE), Valencia, pp. 21-30, 2017.

[18] Park, J-S. Seraku, T. Kiaer, J. "Issues in Defining/Extracting Collocations in Japanese and Korean: Empirical Implications for Building A Collocation Database", Heliyon 2 e00189, 2016.

[19] Das, B. "Extracting Collocations from Bengali Text Corpus", 2nd International Conference on Computer, communication, Control and Information Technology, Kuching, Vol. 4, pp. 325-329, 2012.

[۲۰] ابراهیم‌زاده، دانیال. ملااحمدی، محمد. یوسفان، احمد. "استخراج باهم‌آیی‌های دوتایی و سه‌تایی از پایگاه داده بزرگ بایگانی روزنامه همشهری"، *دومین کنفرانس ملی محاسبات توزیعی و پردازش داده‌های بزرگ*، دانشگاه شهید مدنی تبریز، تبریز، ۱۳۹۵.

[۲۱] مکی، مهدی. "دسته‌بندی موضوعی مطالب وب‌گاه رادیو انگلیسی صدا و سیمای جمهوری اسلامی ایران با الگوریتم‌های شبکه بیزین، Kstar و درخت تصمیم J48"، گزارش پایان‌نامه کارشناسی، دانشکده برق و کامپیوتر، دانشگاه کاشان، کاشان، ۱۳۹۰.

[22] Hyland, K. "As Can Be Seen: Lexical Bundles and Disciplinary Variation", English for Specific Purposes, Vol 27, pp. 4-21, 2008.

[۲۳] پناهی، ثریا. "فرآیند باهم‌آیی و ترکیبات باهمایند در زبان فارسی"، *نامه فرهنگستان*، شماره ۳، دوره ۵، صفحه‌های ۱۹۹-۲۱۱، ۱۳۸۱.

[24] Smadja, F. McKeown, K, R. "Translating Collocations for Bilingual", Journal Computational Linguistics, MIT Press Cambridge, MA, USA, 1996.

[25] Wermter, J. Hahn, U. “Collocation Extraction Based on Modifiability Statistics”, Proceeding COLING '04 Proceedings of the 20th International Conference on Computational Linguistics Association for Computational Linguistics Stroudsburg, PA, USA, 2004.

[۲۶] احسان، نوا. فیلی، هشام. "بررسی تاثیرات ریشه‌یابی در بازیابی اطلاعات در زبان فارسی"، *نشریه پردازش علائم و داده‌ها*، شماره ۱، صفحه‌های ۱۷-۲۴، ۱۳۹۰.

[27] Porter, M, F. “An Algorithm for Suffix Stripping”, Program, Vol. 14, No. 3, pp. 130-137, 1980.

[28] Everta, S. Krennb, B. “Using Small Random Samples for The Manual Evaluation of Statistical Association Measures”, Computer Speech & Language, Vol 19, pp. 450-466, 2005.

[29] Antoch, J. Prchal, L. Sarda, L. “Combining Association Measures for Collocation Extraction Using Clustering of Receiver Operating Characteristic Curves”, Journal of Classification, Springer-Verlag, Vol 30, pp. 100-123, 2013.

[30] Pecina, P. “Lexical Association Measures and Collocation Extraction”, Language Resources and Evaluation, Springer Netherlands, Vol 44, pp. 137-158, 2010.