استخراج مفاهیم کلیدی با استفاده از شبکه قاب و زنجیره مفاهیم

سودابه محمدی و کامبیز بدیع

چکیده: طی سال‌های اخیر، رویکردهای متنوعی جهت استخراج خودکار کلمات و یا عبارات کلیدی ارائه شده است اما رویکردهای اندکی برای استخراج مفاهیم/ نکات کلیدی به طور خودکار وجود دارد که اغلب آنها نیز مبتنی بر متدهای آماری هستند. استخراج مفاهیم کلیدی فرایند شناسایی عباراتی است که بیانگر مفهوم اصلی متن هستند. در این مقاله رویکرد جدیدی جهت استخراج مفاهیم کلیدی با استفاده از شبکه قاب پیشنهاد شده که مبتنی بر پردازش زبان طبیعی است. در این رویکرد، تجزیه معنایی متن اصلی با استفاده از شبکه قاب صورت می‌گیرد و زنجیره‌های مفاهیم ساخته می‌شوند. به هر مفهوم بردار امتیازی متشکل از چهار امتیاز که سه تای آنها مبتنی بر زنجیره‌های مفاهیم هستند، نسبت داده می‌شود. در نهایت مفاهیمی که امتیاز آنها بیش از حد آستانه است، به عنوان مفاهیم کلیدی استخراج می‌شوند. سه حد آستانه متفاوت در این پژوهش مورد استفاده قرار گرفته و در نهایت با یکدیگر مقایسه می‌شوند. برای ارزیابی سیستم پیشنهادی از خبره استفاده می‌شود و معیارهای دقت و یادآوری بررسی می‌شوند. کاربرد مفاهیم کلیدی در مسایلی نظیر شاخص‌گذاری متون الکترونیکی، ساخت کتابخانه‌های دیجیتال، خلاصه‌سازی متون، موتورهای جستجو، خوشه‌بندی، دسته‌بندی و ... است.

کلیدواژه: **استخراج مفاهیم کلیدی، تجزیه‌گر معنایی، زنجیره مفاهیم، شبکه قاب.**

۱- مقدمه

در سال‌های اخیر در نتیجه رشد سریع شبکه و اینترنت، انسان‌ها با حجم عظیم داده به شکل متن، تصویر، صوت و ویدئو مواجه شده‌اند. داده‌های متنی یکی از پرکاربردترین انواع داده است که در کتاب‌ها، مقالات، صفحات وب، ایمیل‌ها، مستندات سازمان‌ها و ... با آنها مواجه می‌شویم. یافتن داده مورد نظر در این میان، کاری دشوار، زمان‌بر و گاهی غیر ممکن به نظر می‌رسد و از طرف دیگر در برخی از کاربردها نیازی به داشتن کل متن نیست. به عنوان مثال در موتورهای جستجو، پرسش با عبارات کلیدی و یا مفاهیم کلیدی موجود در متون مورد جستجو، تطبیق داده می‌شود. به عنوان مثالی دیگر می‌توان به مشابهت بین دو متن اشاره کرد. اگر عبارات کلیدی و یا مفاهیم کلیدی موجود در دو متن بر یکدیگر منطبق شوند، تطبیق بیشتر نشان‌دهنده مشابهت بیشتر است. عبارات کلیدی و یا مفاهیم کلیدی در کاربردهای دیگری نظیر خلاصه‌سازی خودکار متون، خوشه‌بندی، دسته‌بندی و غیره نیز مفید هستند. مثال‌هایی از این دست، لزوم ارائه رویکردهای استخراج عبارات کلیدی و یا مفاهیم کلیدی از متون را نشان می‌دهند. بنابراین این موضوع، یکی از حوزه‌های بااهمیت پژوهشی است و پژوهشگران متعددی در این زمینه مشغول فعالیت هستند که منجر به ارائه رویکردهای مفید و کارایی جهت تحقق این هدف گردیده است.

عبارات کلیدی و یا کلیدواژه‌ها، موضوعات اصلی را که در متن بحث شده است نشان می‌دهند. آنها مجموعه‌ای از واژگان هستند که در متن اصلی وجود دارند و در حقیقت خلاصه‌ای از متن اصلی را نمایش می‌دهند. از طرف دیگر، مفاهیم/نکات کلیدی موجود در متن، تعریفی مشابه عبارات کلیدی دارند با این تفاوت که نشان‌دهنده برخی مفاهیم و معانی اساسی موجود در متن هستند که از لغاتی تشکیل شده‌اند که لزوماً در متن اصلی وجود ندارند. به عنوان مثال متن زیر را که برگرفته از روزنامه همشهری در خصوص جنگ جهانی دوم است در نظر بگیرید:

جنگ جهانی دوم، دومین جنگ فراگیر است که از سپتامبر ۱۹۳۹ آغاز شد و در اوت ۱۹۴۵ پایان یافت. این جنگ علاوه بر اروپا در بخش‌های گسترده‌ای از قاره آسیا و آفریقا تأثیرات مخرب عمده‌ای بر جای گذاشت و کشورهای اسلامی از جمله ایران را درگیر خود کرد. علل اصلی جنگ جهانی دوم عبارت بود از اشتباهات عهدنامه ورسای (۷ مه ۱۹۱۹) که ظاهراً به جنگ جهانی اول پایان داد، همچنین پیامدهای بحران اقتصادی سال ۱۹۲۹ و از همه مهم‌تر رقابت سیاسی فاشیسم و دموکراسی‌های غربی و مارکسیسم. عامل اخیر چنان مؤثر بود که نبرد میان کشورهای درگیر به شکل بی‌سابقه‌ای، عموم مردم را به قلمرو جنگ کشاند به طوری که در پایان جنگ تعداد کشته‌شدگان نظامی و غیر نظامی تقریباً با هم برابری می‌کرد. این جنگ که بین دو بلوک متحدین (آلمان، ایتالیا و ژاپن) و متفقین (انگلیس و فرانسه و آمریکا و شوروی) درگرفت، به لحاظ گستردگی جغرافیایی و قدرت تخریب منابع انسانی و طبیعی، بی‌همتا بوده است (عبدالرضا هوشنگ مهدوی، همشهری آنلاین، ۱۶ شهریور ۱۳۸۷).

آنچه که می‌توان به عنوان مفاهیم کلیدی از متن فوق استخراج کرد مواردی نظیر، قحطی، ناامنی در جهان، خشونت، کشتار انسان‌های بی‌گناه و ... می‌باشند که این عبارات اگرچه عیناً در متن دیده نمی‌شوند اما می‌توان آنها را از متن درک کرد.

در این پژوهش سیستم خودکاری ارائه شده که توانایی استخراج مفاهیم کلیدی از متون انگلیسی را دارد. سرعت و دقت این سیستم در مقایسه با انسان مورد توجه است که هر دو معیار در فصل چهارم مورد ارزیابی قرار می‌گیرند. رویکرد پیشنهادی شامل چندین گام است که به طور خلاصه در ادامه شرح داده می‌شوند.

گام اول در فرایند استخراج مفاهیم کلیدی، پیش‌پردازش است که شامل دو بخش قطعه‌بندی و سپس تفسیر معنایی متن ورودی است که در آن مفهوم هر واژه، شناسایی می‌شود. با توجه به این که هر واژه ممکن است دارای چندین معنی باشد، یافتن نقش معنایی آن در جمله در این مرحله انجام می‌پذیرد. شبکه قاب[1] یک ابزار بسیار قدرتمند برای رسیدن به این هدف است. این شبکه یک پایگاه داده لغوی در زبان انگلیسی است که قابلیت خوانایی توسط ماشین و انسان را دارد و مبتنی بر تفسیر مثال‌هایی از چگونگی به کارگیری واژه‌ها در متن واقعی است. بیش از

این مقاله در تاریخ ۳۱ خرداد ماه ۱۳۹۵ دریافت و در تاریخ ۱۲ آذر ماه ۱۳۹۵ بازنگری شد.

سودابه محمدی، گروه مهندسی کامپیوتر، دانشکده فناوری اطلاعات، دانشگاه صنعتی کرمانشاه، کرمانشاه، (email: su.mohamadii@kut.ac.ir).

کامبیز بدیع، مرکز تحقیقات مخابرات ایران، تهران، (email: k_badie@itrc.ac.ir).

1. Framenet

۱۷۰۰۰۰ جمله تفسیرشده به صورت دستی در شبکه قاب وجود دارند که یک دادگان آموزش‌دیده منحصر به فرد را برای برچسب‌گذاری نقش‌های معنایی فراهم می‌کنند [۱]. آنچه که به عنوان خروجی این گام به دست می‌آید، متن برچسب‌گذاری شده‌ای است که در آن به هر واژه، مفهوم آن واژه نسبت داده شده است.

گام بعدی، ساخت زنجیره‌های مفاهیم و سپس امتیازدهی به آنهاست. زنجیره مفاهیم، مفهومی مشابه زنجیره واژگان دارد با این تفاوت که برای ساخت زنجیره واژگان از واژه‌های موجود در متن استفاده می‌شود حال آن که زنجیره مفاهیم با استفاده از مفاهیم (قاب‌های) موجود در متن ساخته می‌شوند. زنجیره مفاهیم دنباله‌ای از مفاهیم است که با یکدیگر ارتباط معنایی دارند و از طریق این روابط به یکدیگر متصل شده‌اند. زنجیره مفاهیم به صورت گرافی نمایش داده می‌شود که گره‌های آن مفاهیم (قاب‌ها) و یال‌های آن نشان‌دهنده ارتباط بین مفاهیم (قاب‌ها) است [۲].

در گام پایانی باید مفاهیمی که اهمیت بیشتری دارند و عناوین اصلی متن ورودی را دربردارند استخراج شوند. برای رسیدن به این هدف مفاهیم باید بر اساس معیارهای مناسبی انتخاب شوند. به همین منظور به هر مفهوم چهار امتیاز نسبت داده می‌شود که سه تای آنها با استفاده از زنجیره‌های مفاهیم به دست آمده‌اند. سپس حدود آستانه‌ای را در نظر می‌گیریم و مفاهیمی که دارای امتیازی بیش از حد آستانه هستند استخراج می‌شوند.

برای ارزیابی سیستم پیشنهادی از مفاهیم کلیدی استخراج‌شده توسط انسان استفاده می‌کنیم.

در بخش دوم مروری بر کارهای مرتبط پیشین خواهیم داشت. در بخش سوم، رویکرد پیشنهادی با ذکر جزئیات شرح داده خواهد شد. این بخش خود شامل ۵ زیربخش می‌باشد. در بخش چهارم سیستم پیشنهادی آزمایش شده و نتایج آزمایشات ارائه شده‌اند. سپس این نتایج با مفاهیم کلیدی استخراج‌شده توسط خبره مقایسه می‌شوند و بخش پایانی نتیجه‌گیری و پیشنهادهای آتی را در بر دارد.

۲- مروری بر رویکردهای پیشین

اگرچه مقالات متعددی در زمینه استخراج خودکار عبارات کلیدی و واژگان کلیدی ارائه شده است اما پژوهش‌های محدودی برای استخراج خودکار مفاهیم/نکات کلیدی وجود دارد. با توجه به این که رویکردهای پیشنهادی برای هر دو مسأله تا حد زیادی مشابه هستند، در این بخش به مطالعه پیشینه هر دو مورد خواهیم پرداخت.

۲-۱ استخراج خودکار واژگان و عبارات کلیدی

عبارت کلیدی در یک متن بیانگر مفهوم اصلی متن است. استخراج عبارت کلیدی می‌تواند در کاربردهای مختلف پردازش زبان طبیعی از جمله فشرده‌سازی جملات [۳]، خلاصه‌سازی [۴] و [۵]، دسته‌بندی متون [۶]، سیستم‌های پرسش/پاسخ و غیره مفید واقع شود. در حال حاضر رویکردهای متفاوتی با هدف استخراج خودکار واژگان و یا عبارات کلیدی وجود دارند [۶] تا [۱۸].

اولین سیستم‌های استخراج عبارات بااهمیت موجود در متون از روش‌های آماری استفاده می‌کردند. این رویکردها به طور عمومی مبتنی بر پیکره‌های زبان‌شناسی و مشخصه‌های آماری مشتق‌شده از پیکره‌ها هستند. مهم‌ترین مزیت آنها این است که مستقل از زبان هستند و یک تکنیک خاص را می‌توان برای چندین زبان به کار برد اما در مقایسه با رویکردهای زبانشناختی از دقت کمی برخوردارند. در سال ۱۹۷۵ سالتن[1] و همکارانش متدی ارائه دادند که کلمات موجود در متن را بر اساس این که تا چه حد قادر هستند متنی را از دیگر متون موجود در مجموعه، متمایز کنند رتبه‌بندی می‌کرد. در این متد به هر کلمه مقداری نسبت داده می‌شد که این مقدار وابسته بود به تنوع در میانگین تفکیک بین متون. کلماتی که موجب ایجاد بیشترین تفکیک می‌شدند در رتبه‌بندی کلمات در رده بالاتری قرار می‌گرفتند [۱۹]. در سال ۱۹۹۵ کوهن[2] رویکرد آماری ارائه داد که عمل اندیس‌گذاری عبارات را به صورت خودکار انجام می‌داد که او این عمل را "های‌لایت[3]" نامید. این رویکرد مختص دامنه و یا زبان خاصی نبود و از هیچ یک از مؤلفه‌های خاص دامنه، ریشه‌یاب‌ها و لیست‌های توقف[4] و ... استفاده نمی‌کرد. کوهن با استفاده از N-gram عمل شاخص‌گذاری را انجام داد که شباهت زیادی به ریشه‌یاب‌ها داشت اما در شکل عمومی‌تر [۲۰]. در سال ۲۰۰۲ اورتنو[5] و همکارانش اثبات کردند که کلمات بااهمیت یک متن، تمایل به خوشه‌بندی شدن دارند و همچنین بیان کردند که انحراف معیار فاصله بین وقوع‌های پی‌درپی یک کلمه، پارامتر مناسبی برای تعیین این خود- جاذبه است [۲۱].

دسته دیگری از رویکردهای استخراج کلمات/ عبارات کلیدی از روش‌های یادگیری ماشین استفاده می‌کنند که این عمل را به دو صورت بانظارت و یا بدون نظارت انجام می‌دهند. روش‌های بانظارت مانند مسایل طبقه‌بندی عمل می‌کنند. دو کلاس در نظر گرفته می‌شود: در کلاس اول، عبارات کلیدی قرار داده می‌شوند و در کلاس دوم عبارات باقیمانده قرار می‌گیرند. در متدهای بانظارت باید از پیکره‌هایی با متون برچسب‌گذاری شده استفاده کرد. تورنی[6] اولین فردی بود که استخراج عبارات کلیدی را به صورت یک مسأله یادگیری بانظارت مطرح کرد [۷]، سیستم جالب توجه دیگری که برای استخراج عبارات کلیدی مطرح‌شده KEA[7] است. اولی از یک الگوریتم ژنتیک و درخت تصمیم استفاده می‌کند و دیگری الگوریتم یادگیری نیو بیزین[8] را مورد آزمایش قرار می‌دهد. تورنی در الگوریتمش از دو مشخصه استفاده می‌کند: موقعیت کلمه در متن و تکرار کلمه. نتایج آزمایش نشان می‌دهند که الگوریتم ژنتیک عبارات کلیدی بهتری را نسبت به الگوریتم درخت تصمیم C۴.۵ استخراج می‌کند. از طرف دیگر، KEA عبارات کلیدی کاندید را با استفاده از روش‌های لغوی شناسایی می‌کند. سپس مقدار مشخصه را برای هر کاندید محاسبه کرده و در نهایت از یک الگوریتم یادگیری ماشین برای پیش‌بینی این که کدام یک از کاندیدها عبارت کلیدی هستند استفاده می‌کند. KEA کمتر از نیمی از عبارات کلیدی نویسنده را می‌یابد.

در رویکردهای بدون نظارت، تابع امتیازدهی که مبتنی بر تکرار و TFxIDF است استفاده می‌شود. ساده‌ترین متد بدون نظارت برای استخراج عبارات کلیدی از TFIDF برای رتبه‌بندی عبارات کلیدی کاندید استفاده می‌کند و کاندیدهای با بیشترین امتیاز را به عنوان عبارت کلیدی انتخاب می‌کند [۸]. TFIDF عبارات کلیدی را تنها بر اساس بسامد آماریشان رتبه‌بندی می‌کند و به این ترتیب عبارات کلیدی با بسامد پایین

1. Salton
2. Cohen
3. Highlight
4. Stop List
5. Ortuno
6. Turney
7. Key Phrase Extraction Automatic
8. Naïve Bayes

را پیشنهاد نمی‌دهد [۹] و [۱۱]. دسته دیگری از متدهای یادگیری بدون نظارت، متدهای مبتنی بر گراف هستند [۲۲]. این متدها ابتدا یک گراف کلمه بر اساس وقوع کلمات در متن می‌سازند و سپس از شیوه‌های رندم واک[1] (مانند pagerank) برای اندازه‌گیری اهمیت کلمه استفاده می‌کنند. بعد از آن کلمات با رتبه بالا به عنوان عبارات کلیدی انتخاب می‌شوند.

اگرچه زنجیره‌های واژگان اغلب در خلاصه‌سازی متون به کار رفته‌اند، در سال ۲۰۰۷ ارکان و سیسکلی [2] [۱۲] اولین کسانی بودند که زنجیره‌های واژگان را برای استخراج کلمات کلیدی پیشنهاد دادند. آنها الگوریتم یادگیری بانظارت را به کار گرفتند و درخت تصمیم C۴.۵ را به عنوان دسته‌بندی کننده در سیستمشان استفاده کردند. آنها معتقد بودند که زنجیره‌های واژگان، پیوستگی واژگانی را در متن نشان می‌دهند و همچنین کلیدهای بااهمیت مفاهیم معنایی متن را به دست می‌دهند. زنجیره‌های واژگان، نقش مهمی در انتخاب واژگان کلیدی دارند. آنها زنجیره‌های مفاهیم را با استفاده از شبکه واژگان[3] ساختند. سیستم آنها به جای عبارات کلیدی، واژه‌های کلیدی را استخراج می‌کرد زیرا در شبکه واژگان بیشتر عبارت‌ها وجود ندارند. آنها آزمایششان را بر روی دو سیستم انجام دادند که سیستم پایه شامل سه مشخصه آماری بود و سیستم دیگر شامل این سه مشخصه به علاوه چهار مشخصه مبتنی بر زنجیره واژگان بود. نتیجه این شد که مشخصه‌های مبتنی بر زنجیره واژگان مقدار دقت را در فرایند استخراج کلمات کلیدی بهبود می‌بخشند. مشخصه‌های مبتنی بر زنجیره واژگان در این رویکرد، بر اعضای زنجیره‌های واژگان تأکید داشت تا خود زنجیره‌ها.

در سال ۲۰۱۲ هوسی[4] و همکارانش مقاله‌ای نوشتند که پنج متد را که که بر روی شش پیکره آزمایش شده بودند مقایسه می‌کرد. متدهای منتخب، رویکردهای مبتنی بر مترادف، بسامد، بسامد وارونه (IDF)، مقدار C[5] و مقدار NC[6] بودند. این متدها در جهت ارزیابی کارایی و کیفیت نتایج، مورد مقایسه قرار گرفتند و نتیجه این شد که بسامد و IDF و رویکرد مبتنی بر مترادف بهترین الگوریتم‌ها هستند. آنها معتقد بودند که نیازمند معیارهای استانداردتری نظیر دقت و صحت برای ارزیابی هستند و همچنین دامنه وسیع‌تری از پیکره‌ها را برای آزمایش سیستمشان نیاز دارند [۲۳].

در سال ۲۰۱۳ [۱۳] سارکار[7] یک رویکرد ترکیبی برای استخراج عبارات کلیدی از متون پزشکی ارائه داد. رویکرد پیشنهادی او ترکیبی از دو متد بود. اولی وزن‌هایی را به عبارت‌های کلیدی کاندید نسبت می‌داد که مبتنی بر ترکیبی از مشخصه‌هایی نظیر موقعیت، بسامد و IDF بودند و دومی وزن‌هایی را به آنها نسبت می‌داد که از شباهت‌های ساختاری و خصوصیتی عبارات کلیدی استفاده می‌کرد. نتایج آزمایش نشان داد که رویکرد ترکیبی پیشنهادی، بهتر از برخی رویکردها نظیر KEA عمل می‌کند. شناسایی عبارات کلیدی کاندید، همچنان به عنوان اولین گام مؤثر در سیستم‌های استخراج عبارات کلیدی در نظر گرفته می‌شود.

هاسن[8] یک بررسی از رویکردهای استخراج عبارات کلیدی خودکار ارائه داد و به آزمایش منابع اصلی خطا در سیستم‌های موجود پرداخت و در خصوص چالش‌های پیش رو بحث کرد. در پایان سه چالش عمده را معرفی کرد: رسیدگی به متون بزرگ، بهبود برنامه ارزیابی و یکی‌کردن دانش پس‌زمینه [۱۴].

در [۲۴] یک چارچوب جدید برای استخراج خودکار عبارات کلیدی از مستنداتی نظیر متن و تصویر ارائه می‌شود که مبتنی بر الگوریتم یادگیری بانظارت و مشخصه‌های متنی و بصری است. الگوریتم پیشنهادی شامل دو بخش استخراج عبارات کلیدی متنی و عبارات کلیدی بصری است. این کار با استفاده از یک مدل استخراج بانظارت انجام می‌شود. نتیجه این رویکرد در مقایسه با حالتی که تنها از مشخصه‌های متنی استفاده می‌شود، بهتر است.

۲-۲ استخراج خودکار مفاهیم کلیدی

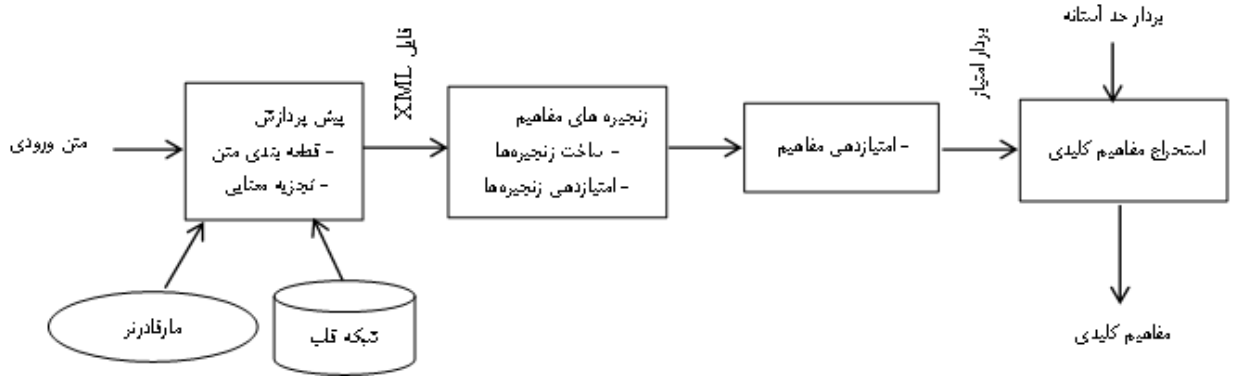
استخراج مفاهیم کلیدی در حل مسایلی نظیر کاربردهای تجارت الکترونیک [۲۵]، پرسش و پاسخ [۲۶]، چکیده‌نویسی، بازیابی اطلاعات و غیره می‌تواند مورد استفاده قرار گیرد.

بنت[9] و همکارانش در سال ۱۹۹۹ اقدام به طراحی یک استخراج‌کننده مفاهیم کردند که می‌توانست به عنوان ماژولی در یک پروژه بزرگ‌تر تحت عنوان "طرح اولیه پروژه اندیس‌گذاری معنایی اطلاعات" مورد استفاده قرار گیرد. در این راستا چهار تجزیه‌گر را مورد بررسی قرار دادند و در نهایت تجزیه‌گر FastNPE را به دلیل سرعت بیشتر و تجزیه‌گر NPtool را به دلیل دقت بیشتر انتخاب کرده و در ماژول استخراج‌کننده مفاهیم، مورد استفاده قرار دادند. نکته جالب توجه این که آنها عبارات اسمی را به عنوان مفاهیم موجود در متن استخراج کردند و تأکید داشتند که استخراج‌کننده مفاهیمی که پیاده‌سازی کرده‌اند در واقع یک استخراج‌کننده عبارت‌های اسمی است. هدف آنها استخراج تمامی مفاهیم موجود در متون، تصاویر و فایل‌های شنیداری بود (و نه تنها مفاهیم کلیدی) [۲۷].

جلفند[10] و همکارانش سیستمی را معرفی کردند که مفاهیم را از یک متن غیر ساختار یافته استخراج می‌کرد. آنها عمل استخراج مفاهیم را با استفاده از شناسایی ارتباط بین کلمات بر پایه پایگاه داده شبکه واژگان انجام دادند و گروه‌هایی از این کلمات تحت عنوان گروه‌های مفهومی را شناسایی کرده که ارتباط معنایی نزدیکی با یکدیگر داشتند. آنها گراف مستقیمی تولید کردند که گراف ارتباط معنایی نام داشت. برای ساخت این گراف از ارتباط بین واژگان استفاده کردند. آنها سپس این سیستم را با دسته‌کننده نیو بیز مقایسه کردند و به این نتیجه رسیدند که دسته‌کننده مبتنی بر گراف ارتباط معنایی به طور قابل توجهی دارای دقتی بیشتر از دسته‌کننده متون نیو بیز است [۲۸].

در سال ۲۰۰۴ رامیرز و متمن[11] الگوریتم ساده‌ای جهت استخراج مفاهیم کلیدی صفحات وب ارائه دادند که کاربرد اصلی آن در موتورهای جستجو بود. اگرچه موتورهای جستجو در پاسخ به سؤال کاربران ممکن است محتویاتی با فرمت‌های pdf، واژه‌پرداز، تصویر، صفحات وب و غیره برگردانند، رامیرز و متمن تنها به بررسی محتویاتی با فرمت زبان نشان‌گذاری فوق متن (HTML) پرداختند. در الگوریتم پیشنهادی ابتدا تمامی برچسب‌های HTML حذف می‌شود و سپس واژگان توقف[12] و

1. Random Walk
2. Ercan and Cicekli
3. Wordnet
4. Hussey
5. C-Value
6. NC-Value
7. Sarkar
8. Hasan
9. Bennett
10. Gelfand
11. Ramirez and Mattmann
12. Stop Words



شکل ۱: معماری سیستم پیشنهادی.

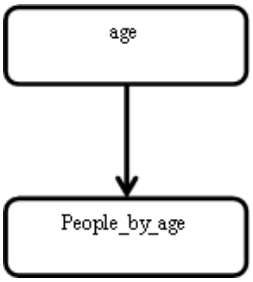
علامت‌های نشانه‌گذاری که قابلیت مفهوم کلیدی‌شدن را ندارند، حذف می‌شوند. در نهایت آنچه که باقی می‌ماند به مجموعه‌ای از توکن‌ها، تبدیل و به هر توکن بنا به بسامد آن، امتیازی نسبت داده می‌شود و در نهایت توکن‌هایی که بسامدی بیش از یک حد آستانه دارند استخراج می‌شوند. چون هر توکن بیانگر یک کلمه است، برای استخراج عبارات به عنوان مفاهیم کلیدی، الگوریتم اقدام به بررسی دامنه محلی توکن‌های استخراج‌شده می‌کند [۲۹].

در سال ۲۰۱۱ محمدی و همکارانش برای اولین بار رویکردی جهت استخراج نکات کلیدی متون ارائه دادند که مبتنی بر هستان‌نگار شبکه قاب بود [۳۰]. آنها در پژوهش خود از زنجیره‌های لغوی استفاده کردند و در ساخت زنجیره‌ها، شبکه قاب را به کار گرفتند. سپس با توجه به معیارهای مختلفی اقدام به امتیازدهی زنجیره‌ها کردند. این معیارها عبارت بودند از معیاری که در [۳۱] توسط بارزیلی[1] و همکارانش استفاده شده بود، تعداد گره‌های زنجیره، مجموع وزن یال‌های زنجیره و حاصل جمع معیارهای دوم و سوم. سپس با در نظر گرفتن یک حد آستانه، تمامی زنجیره‌هایی که وزنشان بیشتر از حد آستانه بود انتخاب شده و گره‌های موجود در آن زنجیره‌ها به عنوان نکات کلیدی استخراج شدند. این سیستم در مقایسه با نکات کلیدی که توسط خبره استخراج شده بود دارای مقدار یادآوری[2] قابل قبولی بود اما میزان دقت[3] پایین بود. علت این نتیجه را می‌توان در این مطلب یافت که تعداد مفاهیمی که در این رویکرد استخراج شدند بسیار زیاد بود و این خود موجب کاهش دقت می‌شد. به نظر می‌رسد که یکی از دلایل رسیدن به این نتیجه این باشد که در این سیستم به جای انتخاب مفاهیم، زنجیره‌ها جهت استخراج انتخاب می‌شدند و زنجیره‌ای که امتیازش مناسب بود، کل مفاهیم موجود در آن به عنوان مفاهیم کلیدی در نظر گرفته می‌شدند. حال آن که ممکن بود مفهوم کم‌اهمیتی به واسطه دیگر مفاهیم بااهمیت موجود در زنجیره، استخراج شود و این ضعف سیستم ارائه‌شده بود. سپس آنها در سال ۲۰۱۲ [۳۲] در تکمیل و اصلاح سیستم پیشنهادی خود، سیستم جدیدی پیشنهاد دادند که به جای استخراج زنجیره‌ها به استخراج مفاهیم می‌پرداخت و معیاری که برای استخراج مفاهیم داشت، انتخاب ۵ مفهوم با امتیاز بیشتر بود زیرا با آزمایش مقدار میانگین به علاوه دو برابر انحراف معیار، به عنوان حد آستانه، به نتایج قابل قبولی نرسیدند. در این سیستم برای هر مفهوم چهار مشخصه در نظر گرفته شده بود که هر یک را به طور مجزا برای انتخاب مفهوم کلیدی استفاده می‌کردند. در نهایت مقدار دقت نسبت به رویکرد پیشین بهبود یافت. در رویکرد پیشنهادی در مقاله پیش رو، اول این که امتیاز هر مفهوم را به صورت برداری با ۴ مؤلفه در نظر گرفته‌ایم. دوم این که سه حد آستانه معرفی شده و نتایج به کارگیری آنها مورد بحث قرار گرفته و در نهایت مناسب‌ترین آنها معرفی شده است زیرا همان طور که پیشتر ذکر شد یکی از ضعف‌های سیستم پیشین نبود حد آستانه مناسب بود. همچنین در پژوهش حاضر، تعداد متون مورد آزمایش و تعداد خبره‌ها افزایش مناسبی داشته که این خود موجب نزدیک‌شدن نتایج به واقعیت خواهد شد.

در سال ۲۰۱۳ متدی جهت استخراج مفاهیم کلیدی ارائه شد که در اولین گام، اقدام به ابهام‌زدایی از کلمات با استفاده از شبکه واژگان می‌کرد. در این راستا گراف مفهومی تولید می‌شود که یال‌های آن با استفاده از روابط holonym و hypernym موجود در شبکه واژگان به دست می‌آیند. این سیستم با مدل کلمات کلیدی TF-IDF مقایسه شده و در ارزیابی سیستم بر مسأله دسته‌بندی، تمرکز شده است. در نهایت نویسندگان این مقاله به این نتیجه رسیدند که اگر در الگوریتمشان از دسته‌کننده نیو بیزین استفاده کنند، در دسته‌بندی متون، نتایج بهتری نسبت به مدل کلمات کلیدی TF-IDF به دست خواهند آورد [۳۳].

یکی از کاربردهای بسیار بااهمیت استخراج مفاهیم، استخراج مفاهیم مربوط به هستان‌شناسی در دامنه‌های خاص می‌باشد. هستان‌شناسی دو بخش دارد: مفاهیم و رابطه‌ها. تشخیص مفاهیم موجود در منابع داده جهت ساخت یک هستان‌شناسی، یکی از کاربردهای سیستم‌های خودکار استخراج مفاهیم کلیدی می‌باشد. در سال ۲۰۱۶ پژوهشی انجام گرفته که اقدام به استخراج مفاهیم هستان‌شناسی از میان چندین متن (تعداد متون محدودیتی ندارد) می‌کند [۳۴] و تمامی متون در یک حوزه خاص قرار دارند (مثلاً گزارش‌های هواشناسی، گزارش‌های ورزشی، گزارش‌های پزشکی و غیره). الگوریتم ارائه‌شده شامل سه مرحله است. در مرحله اول عمل پیش‌پردازش متون انجام می‌شود که شامل قطعه‌بندی کردن کلمات است. سپس بر اساس الگوریتم N-gram مجموعه‌ای از عبارات کاندید تولید می‌شوند و در پایان با استفاده از قوانین آماری و زبان‌شناختی اقدام به استخراج مفاهیم می‌کند. دو معیار اطلاعات متقابل و بسامد مستندات به عنوان قوانین آماری برای استخراج مفاهیم از بین عبارات کاندید استفاده می‌شوند. با توجه به این که اطلاعات متقابل سعی در انتخاب مفاهیم با بسامد پایین دارد، معیار بسامد مستندات در کنار آن این نقص را برطرف می‌کند. حذف واژگان توقف نیز به عنوان قوانین زبان‌شناختی مورد استفاده قرار می‌گیرد. اگرچه در اینجا مفاهیم کلیدی مد نظر نیست اما روال استخراج مفاهیم دقیقاً مشابه روال استخراج مفاهیم کلیدی است.

1. Barzilay
2. Recall
3. Precision



شکل ۲: مثالی از یک رابطه قوی بین دو قاب age و people-by-age.

۳- رویکرد پیشنهادی

معماری سیستم پیشنهادی در شکل ۱ نشان داده شده است. ورودی این سیستم متن اصلی و خروجی آن مجموعه‌ای از مفاهیم کلیدی است. همان طور که در تصویر مشاهده می‌کنید این سیستم دارای چهار مرحله اصلی است که خروجی هر مرحله، ورودی مرحله بعدی می‌باشد. در ادامه این مراحل با جزئیات شرح داده می‌شوند.

۳-۱ پیش‌پردازش متن

ورودی این مرحله متن اصلی است و خروجی آن یک فایل xml است که نشان‌دهنده تجزیه معنایی متن ورودی می‌باشد. این مرحله خود شامل دو زیربخش است، ابتدا متن ورودی قطعه‌بندی می‌شود و سپس عمل برچسب‌گذاری نقش‌های معنایی صورت می‌پذیرد.

۳-۱-۱ قطعه‌بندی متن ورودی

در اولین مرحله، متن ورودی باید با استفاده از روش‌های قطعه‌بندی، قطعه‌بندی شود. قطعه‌بندی متن، روند تقسیم متن ورودی به واحدهای معناداری نظیر کلمه، جمله و یا موضوع است. قطعه‌کننده متن ابزاری برای قطعه‌بندی متن است که سعی دارد متن را به قطعات معنادار تقسیم کند. ابزارهای مختلفی برای این هدف وجود دارند که هر یک از آنها از روش‌های قطعه‌بندی خاص خود استفاده می‌کنند. در این مقاله از قطعه‌کننده مارفادرنر[1] استفاده شده است. این ابزار از دو متد قطعه‌بندی خطی [۳۵] و [۳۶] استفاده می‌کند. دلیل اصلی قطعه‌بندی کردن متن در این رویکرد این است که از ساخت زنجیره‌های خیلی بزرگ پیشگیری شود. اگر زنجیره‌های مفاهیم با استفاده از مفاهیم موجود در کل متن ساخته شوند، اندازه زنجیره‌ها بسیار بزرگ می‌شود و حجم زیادی برای ذخیره این زنجیره‌ها نیاز است. از طرفی تولید این زنجیره‌ها بسیار زمان‌بر و پیچیده خواهد شد. بنابراین متن به قطعات معنایی کوچک‌تر تقسیم شده و برای هر قطعه از مفاهیم موجود در همان قطعه برای ساخت زنجیره‌هایش استفاده شده است.

۳-۱-۲ تجزیه معنایی

تجزیه معنایی متن به معنای یافتن ساختار معنایی آن می‌باشد. استخراج نقش‌های معنایی یکی از گام‌های اصلی در بازنمایی معنی متن است. نقش‌های معنایی، ارتباط معنایی بین فعل و آرگومان‌های آن را در جمله مشخص می‌کنند. مشخص‌کردن ساختار معنایی متن یکی از عملیات‌های کلیدی و اصلی در کاربردهای پردازش زبان طبیعی است (کاربردهایی نظیر استخراج اطلاعات، خلاصه‌سازی، پرسش/ پاسخ، شباهت معنایی، ترجمه ماشینی و ...) [۳۷]. در سال‌های اخیر تعداد زیادی از پژوهشگران این مسأله را به عنوان یک مسأله برچسب‌گذاری مطرح کرده‌اند و از تکنیک‌های یادگیری ماشین برای ساخت چنین سیستم‌هایی استفاده می‌کنند. برخی از این سیستم‌ها عمل آموزش را با استفاده از تفسیر شبکه قاب انجام می‌دهند و به طور خودکار تفسیرهایی برای متون ارائه می‌دهند. این روند را می‌توان برچسب‌گذاری خودکار نقش‌های معنایی نامید. در حال حاضر سه سیستم شناخته‌شده وجود دارند که عمل برچسب‌گذاری خودکار نقش‌های معنایی را با استفاده از شبکه قاب انجام می‌دهند. این سیستم‌ها عبارتند از شلمنزر[2] [۳۸]، LTH [۳۹] و [۴۰] و سمافور[3] [۴۱] و [۴۲]. همچنین پروژه شبکه قاب خود دارای تعدادی متن کامل تفسیر شده است که این متون به صورت دستی و با استفاده از این دادگان تفسیر شده‌اند در این مقاله از این متون کامل، جهت ارزیابی سیستم استفاده می‌شود.

۳-۲ ساخت زنجیره‌های مفاهیم

همان طور که در شکل ۱ مشاهده شد، مرحله بعدی در روند استخراج مفاهیم کلیدی، ساخت زنجیره‌های مفاهیم است. یک زنجیره مستقل از ساختار گرامری متن است. یک زنجیره مفاهیم دنباله‌ای از مفاهیم مرتبط موجود در متن اصلی است. در این پژوهش، زنجیره‌های مفاهیم با استفاده از قاب‌هایی که به واحدهای لغوی نسبت داده شده‌اند، ساخته می‌شوند. ارتباط بین مفاهیم در زنجیره‌ها بر اساس ارتباطات قاب موجود در شبکه قاب می‌باشد. برای ساخت این زنجیره‌ها از الگوی ساخت زنجیره‌های لغوی که در [۲۸] و [۲۹] معرفی شده‌اند، استفاده می‌شود. ساخت این زنجیره‌ها در سه مرحله انجام می‌پذیرد:

الف) انتخاب ترم‌های کاندید

در این مرحله، قاب‌های موجود در متن را به عنوان ترم‌های کاندید در نظر می‌گیریم. در واقع، قاب بیانگر مفهوم یک واژه در موقعیت خاصی از متن است. به همین دلیل زنجیره‌هایی را که از قاب‌ها تشکیل شده‌اند تحت عنوان زنجیره‌های مفاهیم نام‌گذاری کرده‌ایم.

ب) یافتن زنجیره مناسب

ارتباط و هماهنگی بین مفاهیم، یکی از مسایل اساسی در ساخت زنجیره‌های مفاهیم است. همان طور که پیشتر اشاره کردیم برای یافتن ارتباطات و فاصله معنایی بین مفاهیم از شبکه قاب استفاده می‌کنیم. سه نوع ارتباط بین مفاهیم تعریف شده است:

ارتباط بسیار قوی: میان یک مفهوم و تکرارهای آن برقرار است.

برای مثال واحدهای لغوی "plan.n" و "program.n" هر دو قاب "project" را فراخوانی می‌کنند. بنابراین اگر این دو واحد لغوی در یک قطعه از متن وجود داشته باشند، باعث تکرار در فراخوانی قاب "project" می‌شوند. همچنین اگر یکی از این دو واحد لغوی خود به صورت مکرر در قطعه‌ای موجود باشد، این نیز منجر به تکرار در فراخوانی قاب مربوطه خواهد شد. فاصله معنایی بین یک قاب و تکرار آن، ده در نظر گرفته شده است.

ارتباط قوی: میان دو قاب که به وسیله یکی از انواع رابطه‌های قاب- با- قاب[4] به یکدیگر متصل شده‌اند برقرار است.

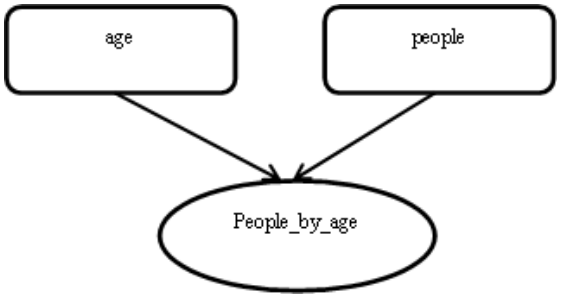
برای مثال دو واحد لغوی "children.n" و "young.a" را در نظر بگیرید که به ترتیب قاب‌های "people_by_age" و "age" را فراخوانی می‌کنند. این دو قاب با استفاده از یک رابطه قاب- با- قاب با یکدیگر در

1. Marphadorner

2. Shalmaneser

3. Semafor

4. Frame-to-Frame



شکل ۳: مثالی از یک ارتباط متوسط بین دو قاب age و people.

ارتباطند و وزن این نوع رابطه هفت در نظر گرفته شده است. تصویر این نوع رابطه را در شکل ۲ می‌توان مشاهده کرد.

لازم به توضیح است که در شبکه قاب، بین دو قاب ممکن است رابطه نامتقارنی وجود داشته باشد که در این رابطه یکی از قاب‌ها که وابستگی کمتری به دیگری دارد به عنوان فوق قاب[1] و دیگری به عنوان زیرقاب[2] شناخته می‌شود. در شکل ۲ قاب age فوق قاب و people_by_age زیرقاب است. در مجموع هشت نوع رابطه قاب- با- قاب در شبکه قاب تعریف شده که توضیحات بیشتر آنها را در [۱] می‌توانید مطالعه کنید.

ارتباط متوسط: میان دو قاب که به وسیله یک قاب دیگر، تحت عنوان قاب واسطه به یکدیگر متصل شده‌اند، برقرار است.

به عنوان مثال، دو واحد لغوی "women.n" و "young.a" که به ترتیب قاب‌های "people" و "age" را فراخوانی می‌کنند در نظر بگیرید. این دو قاب به وسیله قاب "people-by-age" به یکدیگر متصل شده‌اند که این قاب را تحت عنوان قاب واسطه در نظر می‌گیریم. تصویر این نوع رابطه در شکل ۳ آمده است. برای این نوع رابطه وزن سه در نظر گرفته شده و لازم به ذکر است که برای این نوع وزن‌دهی به پژوهش [۳۱] استناد شده است.

هر قاب کاندید به ترتیب حضورش در قطعه مورد نظر انتخاب شده تا زنجیره مناسب آن پیدا شود. فرض کنید تاکنون n زنجیره تشکیل شده و قاب x به عنوان قاب کاندید در نظر گرفته شده تا در زنجیره مناسبش قرار بگیرد. در مرحله یافتن زنجیره مناسب برای هر قاب کاندید سه حالت مختلف ممکن است رخ دهد. اگر x با هیچ یک از قاب‌های موجود در n زنجیره قبلی ارتباط نداشته باشد، یک زنجیره جدید تشکیل شده و x به عنوان اولین عنصر در آن درج می‌شود. اگر تنها یکی از n زنجیره قبلی برای x مناسب باشد آن گاه x به آن اضافه می‌شود. در حالت سوم اگر بیش از یک زنجیره مناسب برای x وجود داشته باشد آن دو زنجیره به واسطه x به یکدیگر متصل شده و زنجیره بزرگ‌تری را تشکیل می‌دهند.

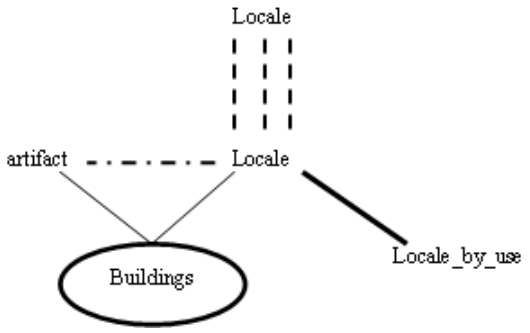
ج) اضافه‌کردن قاب به زنجیره

هرگاه زنجیره مناسب قاب یافت شد، قاب به زنجیره اضافه می‌گردد به این ترتیب که بین قاب جدید و دیگر قاب‌های موجود در زنجیره که با آن ارتباط دارند، یال(هایی) متصل می‌شوند و زنجیره به روز می‌شود.

۳-۳ امتیازدهی به زنجیره‌های مفاهیم

در این مرحله زنجیره‌ها بر اساس معیارهای مختلف امتیازدهی می‌شوند. معیارهای متفاوتی جهت امتیازدهی زنجیره‌ها وجود دارد از جمله تعداد عناصر (گره‌های) یک زنجیره، حاصل جمع وزن یال‌های موجود در زنجیره، توزیع عناصر زنجیره در متن، پوشش محدوده متن توسط زنجیره،

1. Super Frame
2. Sub Frame



شکل ۴: محاسبه مشخصه ۴ مربوط به مفهوم Locale.

چگالی، توپولوژی گراف، تعداد تکرارهای عناصر زنجیره و ... [۳۱]. در پژوهش‌های پیشین، تعداد عناصر زنجیره به عنوان بهترین معیار سنجش قدرت زنجیره‌ها شناخته شده است. در این پژوهش، مجموع وزن یال‌های زنجیره نیز به عنوان معیاری دیگر برای امتیازدهی زنجیره‌ها به کار رفته است. بنابراین هر زنجیره دارای دو امتیاز است:

امتیاز ۱: تعداد عناصر (گره‌های) زنجیره

امتیاز ۲: مجموع فاصله‌های معنایی موجود در زنجیره (مجموع وزن یال‌های زنجیره)

۴-۳ امتیازدهی به مفاهیم

همان طور که پیشتر ذکر شد در این مقاله بر خلاف [۳۰] در روند استخراج به جای استفاده از امتیاز زنجیره‌ها از امتیاز مفاهیم استفاده می‌شود، با این انگیزه که دقت سیستم بهبود یابد. بنابراین چهار مشخصه متفاوت برای امتیازدهی به مفاهیم استفاده می‌گردد که این مشخصه‌ها عبارتند از:

مشخصه ۱ (بسامد مفهوم): مقدار این مشخصه برابر است با تعداد تکرارهای یک مفهوم در متن.

مشخصه ۲: بیشترین امتیاز ۱ مربوط به زنجیره‌ای که مفهوم مورد نظر عضوی از آن زنجیره است.

یک مفهوم ممکن است عضو زنجیره‌های مختلفی از متن باشد که هر یک از این زنجیره‌ها دارای یک امتیاز ۱ می‌باشند. از بین این زنجیره‌ها هر کدام که بیشترین امتیاز ۱ را دارد، امتیاز ۱ آن به عنوان مشخصه شماره ۲ مفهوم در نظر گرفته شده است.

مشخصه ۳: بیشترین امتیاز ۲ مربوط به زنجیره‌ای که مفهوم مورد نظر عضوی از آن زنجیره است.

این مشخصه مشابه مشخصه ۲ است با این تفاوت که در آن به جای امتیاز ۱ زنجیره‌ها، امتیاز ۲ آنها در نظر گرفته شده است.

مشخصه ۴ (بیشترین وزن رابطه‌های مفهوم): برای محاسبه این مشخصه مجموع وزن یال‌هایی را که به مفهوم متصل هستند در زنجیره‌های مختلف، محاسبه کرده و بیشترین آنها را به عنوان مشخصه ۴ برای مفهوم در نظر می‌گیریم.

به عنوان مثال در شکل ۴، امتیاز مفهوم Locale برابر ۴۱ است به این ترتیب که این مفهوم به دلیل سه بار تکرار شدن در قطعه مورد نظر، دارای ۳ رابطه بسیار قوی با وزن ۱۰ است (که با خط‌چین مشخص شده است). همچنین یک رابطه قوی با مفهوم Locale_by_use دارد که وزن آن ۷ می‌باشد (که با خط پررنگ مشخص شده است). یک رابطه متوسط با مفهوم Artifact دارد که دارای وزن ۴ است (این رابطه با خطنقطه مشخص شده و مفهوم Buildings به عنوان مفهوم واسطه در این رابطه است). مجموع وزن این رابطه‌ها برابر ۴۱ است.

جدول ۱: میانگین یادآوری مفاهیم کلیدی استخراج‌شده توسط سیستم پیشنهادی در مقایسه با خبره با استفاده از سه حد آستانه.

میانگین + دو برابر انحراف معیار	میانگین + انحراف معیار	میانگین
۰/۴۳	۰/۵۷	۰/۴۴

جدول ۲: میانگین دقت مفاهیم کلیدی استخراج‌شده توسط سیستم پیشنهادی در مقایسه با خبره با استفاده از سه حد آستانه.

میانگین + دوبرابر انحراف معیار	میانگین + انحراف معیار	میانگین
۰/۱۷	۰/۲۱	۰/۲۷

۳-۵ استخراج مفاهیم کلیدی

این مرحله، مرحله پایانی است. در این مرحله تعدادی از مفاهیم به عنوان مفهوم کلیدی استخراج می‌شوند. برای رسیدن به این هدف سه حد آستانه در نظر گرفته شده است. اگر امتیاز یک مفهوم بیشتر از حد آستانه باشد به عنوان مفهوم کلیدی استخراج می‌شود و در صورتی که امتیاز مفهوم کمتر از حد آستانه باشد، مفهوم کلیدی نخواهد بود. سه حد آستانه با توجه به فرمول

$$\text{Score(concept)} > \text{Average(scores)} + C \times \text{StandardDeviation(Scores)} \quad (۱)$$

که در [۳۰] و [۳۱] مطرح شده است در نظر گرفته شده‌اند. در اینجا ثابت C وابسته به تعداد کل مفاهیم موجود در متن و یا تعداد مفاهیمی است که سیستم می‌تواند استخراج کند. در این مقاله محدودیتی بر روی تعداد مفاهیمی که سیستم می‌تواند استخراج کند قرار نداده‌ایم. به C مقادیر صفر، یک و دو می‌دهیم و سه حد آستانه زیر به دست می‌آیند.

حد آستانه ۱: میانگین امتیاز تمامی مفاهیم موجود در متن

حد آستانه ۲: میانگین امتیاز تمامی مفاهیم موجود در متن + انحراف معیار امتیاز مفاهیم موجود در متن

حد آستانه ۳: میانگین امتیاز مفاهیم موجود در متن + دو برابر انحراف معیار امتیاز مفاهیم موجود در متن

این سه حد آستانه با توجه به نمودار توزیع نرمال انتخاب شده‌اند. انتظار می‌رود که ۵۰ درصد مفاهیم دارای امتیازی بیش از میانگین (حد آستانه ۱)، ۱۵/۹ درصد آنها دارای امتیازی بیش از میانگین به علاوه یک انحراف معیار (حد آستانه ۲) و حدود ۲/۳ درصد آنها دارای امتیازی بیش از میانگین به علاوه دو برابر انحراف معیار (حد آستانه ۳) باشند.

برای هر مفهوم یک بردار با چهار مؤلفه در نظر گرفته می‌شود که این چهار مؤلفه عبارتند از امتیازهای یک تا چهار که بر اساس مشخصه‌های یک تا چهار به هر مفهوم نسبت داده می‌شوند. مفاهیمی به عنوان مفهوم کلیدی استخراج می‌شوند که بردار مربوط به آنها، بزرگ‌تر از بردار حد آستانه باشد.

۴- نتایج آزمایش‌ها و ارزیابی آنها

شبکه قاب شامل چندین متن تفسیرشده است که نشان می‌دهد چگونه معانی قاب می‌توانند در درک و فهم متن کمک کنند. تفسیر این متون به صورت دستی و با استفاده از هستان‌نگار شبکه قاب صورت گرفته و بنابراین می‌توان برای ارزیابی سیستم پیشنهادی از آنها استفاده کرد. این متون از پیکره‌های مختلفی انتخاب شده‌اند. در این پژوهش از ده متن تفسیرشده شبکه قاب برای ارزیابی سیستم استفاده می‌شود که هشت تای آنها مربوط به پیکره LU[1] و دو تای دیگر از پیکره ملی آمریکایی[2] می‌باشند. در پیکره LU مجموعه‌ای از اسناد متنوع شامل مکالمات تلفنی، پست‌های الکترونیکی و ترجمه‌هایی از متون عربی وجود دارد.

در ارزیابی سیستم جمعاً ۵۲۵ مفهوم در ده متن وجود داشت یعنی به طور میانگین در هر متن حدود ۵۲ مفهوم شناسایی شد که پس از اجرای الگوریتم استخراج، تعدادی به عنوان مفهوم کلیدی و باقی به عنوان مفاهیم غیر کلیدی شناخته شدند.

تورنی معتقد است که هر عبارتی به صورت بالقوه می‌تواند عبارت کلیدی باشد اما تنها آن عباراتی، عبارت کلیدی هستند که با عبارات کلیدی انتساب‌شده توسط انسان، مطابقت داشته باشند [۷]. بنابراین برای آزمایش سیستم، ابتدا با استفاده از سیستم پیشنهادی مفاهیم کلیدی این متون شناسایی شده و این مفاهیم با مفاهیمی که توسط خبره استخراج شده‌اند مقایسه می‌شوند. برای هر متن ۵ مورد استخراج مفاهیم کلیدی به صورت دستی توسط پنج نفر انجام می‌شود. در کل تعداد ۳۰ نفر به عنوان خبره برای ارزیابی این سیستم همکاری کردند که این افراد همگی دانشجوی رشته مهندسی کامپیوتر دانشگاه صنعتی کرمانشاه بوده و تسلط آنها به زبان انگلیسی در حد متوسط است. محدودیتی در تعداد مفاهیمی که باید از هر متن استخراج شوند وجود ندارد.

ما در این مقاله برای ارزیابی توافق میان افراد در استخراج نکات کلیدی از مقیاس توافق درصد که در [۴۳] تعریف شده، استفاده کرده‌ایم. در این آزمایش توافق درصد برابر ۴۷٪ می‌باشد. بارزیلی در آزمایش سیستمش میزان توافق درصد ۹۳٪ را بین پنج نفر به دست آورد که این مقدار در مقایسه با دیگر پژوهش‌هایی که در این زمینه انجام شده بود (با معیارهای توافق ۲۵٪، ۴۶٪ و غیره) غیر منتظره بود [۳۱].

دو معیار کارایی که در ارزیابی این سیستم استفاده شده است عبارتند از یادآوری و دقت.

نتایج حاصل از مقایسه سیستم خودکار و خبره در جداول ۱ و ۲ قابل مشاهده است. همان طور که در جداول فوق می‌توان مشاهده کرد، مقدار یادآوری به طور کلی بیشتر از دقت سیستم است. همچنین بیشترین مقدار یادآوری به دست آمده مربوط به حد آستانه ۲ و بیشترین مقدار دقت مربوط به حد آستانه ۱ است. با توجه به این که در توزیع نرمال تنها ۲/۳ درصد از مفاهیم در فاصله به علاوه دو انحراف معیار قرار می‌گیرند، و بنا بر آنچه که پیشتر گفتیم در هر متن به طور میانگین حدود ۵۲ مفهوم وجود دارد، بنابراین با استفاده از این حد آستانه از هر متن تقریباً یک مفهوم توسط سیستم استخراج می‌شود. به همین دلیل این حد آستانه مناسب به نظر نمی‌رسد. اما حد آستانه به علاوه یک انحراف معیار شامل ۱۵/۹ درصد مفاهیم است که با توجه به میانگین فوق از هر متن حدود ۷ مفهوم در این بازه قرار دارد.

شاید بتوان این را به عنوان توضیحی در پایین‌بودن مقدار دقت بیان کرد که استخراج مفاهیم کلیدی یک فرایند دقیق نیست زیرا انسان نیز در این عمل ممکن است درگیر تناقض و یا اشتباه شود.

در تحلیل پیچیدگی زمانی سیستم باید پیچیدگی زمانی سه مرحله کلی تجزیه معنایی، ساخت زنجیره‌های مفاهیم و استخراج مفاهیم کلیدی را در نظر گرفت. ساخت زنجیره‌های مفاهیم در زمانی خطی انجام می‌گیرد [۴۴]. برای استخراج مفاهیم کلیدی نیاز است مقدار بردار امتیاز هر مفهوم با بردار حد آستانه مقایسه شود که پیچیدگی زمانی این مرحله نیز خطی

1. LU-Corpus
2. American National Corpus

است. بنابراین پیچیدگی زمانی الگوریتم پیشنهادی وابسته به پیچیدگی زمانی تجزیه‌گر معنایی است.

۵- نتیجه‌گیری و پیشنهادهای آتی

تاکنون رویکردهای متعددی با هدف خلاصه‌سازی متون و استخراج عبارات کلیدی از متون ارائه شده‌اند اما برای استخراج مفاهیم کلیدی، رویکردهای اندکی وجود دارد. در این مقاله، رویکردی جهت استخراج مفاهیم کلیدی ارائه شده است. برای رسیدن به این هدف، زنجیره‌هایی تحت عنوان زنجیره‌های مفاهیم استفاده می‌شوند که ساخت آنها مبتنی بر شبکه قاب است. اگرچه زنجیره‌های لغوی در دامنه گسترده‌ای از کاربردها از جمله خلاصه‌سازی متون و استخراج کلمات کلیدی استفاده شده‌اند اما در این مقاله برای اولین بار، این ابزار تحت عنوان زنجیره مفاهیم استفاده می‌شود. چهار مشخصه برای هر مفهوم در نظر گرفته شده که بر اساس هر یک از آنها امتیازی به مفهوم نسبت داده می‌شود. سه تا از مشخصه‌ها مبتنی بر زنجیره‌های مفاهیم هستند. برای هر مفهوم، برداری متشکل از چهار امتیاز در نظر گرفته می‌شود و مقایسه با حد آستانه جهت استخراج مفهوم با استفاده از این بردار چهار مؤلفه‌ای انجام می‌شود. نتایج نشان می‌دهند که استفاده از حد آستانه "میانگین + یک انحراف معیار" دارای یادآوری بیشتری است در حالی که حد آستانه "میانگین"، دقت بیشتری را به دنبال دارد.

با توجه به این که تاکنون تنها ۱۰۰۰۰ ورودی برای شبکه قاب تعریف شده و این پروژه در حال توسعه و گسترش است، این امکان وجود دارد که برخی از مفاهیم موجود در متن هنوز در شبکه قاب تعریف نشده باشند و نادیده گرفته شوند. بنابراین در کارهای آتی می‌توان برای ساخت زنجیره‌های مفاهیم به طور هم‌زمان از شبکه قاب و شبکه واژگان (با ۱۱۷۰۰۰ مجموعه مترادف) استفاده کرد.

مراجع

[1] —, *General Release Note 1.5.* https://framenet.icsi.berkeley.edu/fndrupal/ Accessed: on 12 Feb. 2016.

[2] T. N. Erekhinskaya and D. I. Moldovan, "Lexical chains on wordnet and extensions," in *Proc. Twenty-sixth Int. Florida Artificial Intelligence Research Society Conf.*, pp. 52-57, May 2013.

[3] F. Boudin and E. Morin, "Keyphrase extraction for N-best reranking in multi-sentence compression," in *Proc. of the NAACL HLT Conf.*, pp. 298-305, Jun. 2013.

[4] J. R. Thomas, S. K. Bharti, and K. S. Babu, "Automatic keyword extraction for text summarization in e-newspapers," in *Proc. ACM Int. Conf. on Informatics and Analytics.*, p. 86, Aug. 2016.

[5] E. D'Avanzo, B. Magnini, and A. Vallin, "Keyphrase extraction for summarization purposes: the LAKE system at DOC-2004," in *Proc. Document Understanding Conf.*, 4 pp., May 2004.

[6] A. Onan, S. Korukoglu, and H. Bulut, "Ensemble of keyword extraction methods and classifiers in text classification," *Expert Syst. Appl.*, vol. 57, pp. 232-247, Sept. 2016.

[7] P. Turney, "Learning algorithms for keyphrase extraction," *Information Retrieval*, vol. 2, no. 4, pp. 303-336, Boston, Sept. 2000.

[8] I. Witten, G. Paynter, and E. Frank, "KEA: practical automatic keyphrase extraction," in *Proc. of the 4th ACM Conf. on Digital Libraries*, pp. 254-255, Aug. 1999.

[9] F. Liu, D. Pennell, and Y. Liu, "Unsupervised approaches for automatic keyword extraction using meeting transcripts," in Proc. of Human Language Technologies: the 2009 Annual Conf. of the North American Chapter of the Association for Computational Linguistics, pp. 620-628, May 2009.

[10] R. Wang, W. Liu, and C. McDonald, "How preprocessing affects unsupervised keyphrase extraction," *Proc. 15th Int. Conference on Computational Linguistics and Intelligent Text Processing , CICLing'014,* pp. 163-176, Kathmandu, Nepal, 6-12 Apr. 2014.

[11] T. Tomokiyo and M. Hurst, "A language model approach to keyphrase extraction," in *Proc. of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, vol. 18, pp. 33-40, Nov. 2003.

[12] G. Ercan and I. Cicekli, "Using lexical chains for keyword extraction," *Information Processing & Management*, vol. 43, no. 6, pp. 1705-1714, Sept. 2007.

[13] K. Sarkar, *A Hybrid Approach to Extract Keyphrases from Medical Documents*, arXiv Prepr. arXiv1303.1441, 2013.

[14] K. Sarkar, M. Nasipuri, and S. Ghose, *A New Approach to Keyphrase Extraction Using Neural Networks*, arXiv Prepr. arXiv1004.3274, 2010.

[15] K. Hasan and V. Ng, "Automatic keyphrase extraction: a survey of the state of the art," in *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 1262-1273, 2014.

[16] S. Kim, O. Medelyan, M. Kan, and T. Baldwin, "Automatic keyphrase extraction from scientific articles," *Lang. Resour. Eval.*, vol. 47, no. 3, pp. 723-742, Sept. 2013.

[17] S. Kim, O. Medelyan, M. Kan, and T. Baldwin, "Semeval-2010 task 5: automatic keyphrase extraction from scientific articles," in *Proc. of the 5th Int. Workshop on Semantic Evaluation, Association for Computational Linguistics*, pp. 21-26, Jul. 2010.

[18] F. Xie, X. Wu, and X. Zhu, "Document-specific keyphrase extraction using sequential patterns with wildcards," in *Proc. IEEE Int. Conf. on Data Mining, ICDM'14*, pp. 1055-1060, Dec. 2014.

[19] G. Salton, C. S. Yang, and C. T. Yu, "A theory of term importance in automatic text analysis," *J. Am. Soc. Inf. Sci.*, vol. 26, no. 1, pp. 33-44, Jan. 1975.

[20] J. D. Cohen, "Highlights: language-and domain-independent automatic indexing terms for abstracting," *J. Am. Soc. Inf. Sci.*, vol. 46, no. 3, p. 162, Apr. 1995.

[21] M. Ortuno, P. Carpena, P. Bernala-Galvan, E. Munoz, and A. M. Somoza, "Keyword detection in natural languages and DNA," *Europhysics Lett.*, vol. 57, no. 5, pp. 759-764, Mar. 2002.

[22] S. Beliga, A. Mestrovic, and S. Martincic-Ipsic, "An overview of graph-based keyword extraction methods and approaches," *Inf. Organ. Sci.*, vol. 39, no. 1, pp. 1-20, Jun. 2015.

[23] R. Hussey, S. Williams, and R. Mitchell, "Automatic keyphrase extraction: a comparison of methods," in *Proc. Int. Conf. on Information Processing and Knowledge Management*, pp. 18-23, Jan. 2012.

[24] Y. HaCohen-Kerner, S. Vrochidis, D. Liparas, A. Moumtzidou, and I. Kompatsiaris, "Keyphrase extraction using textual and visual features," in *Proc. 25th Int. Conf. Comput. Linguist.*, pp. 121-123, Aug. 2014.

[25] Y. Zhang, R. Mukherjee, and B. Soetarman, "Concept extraction and e-commerce applications," *Electronic Commerce Research and Applications*, vol. 12, no. 4, pp. 289-296, Aug. 2013.

[26] D. Glinos, *Syntax-Based Concept Extraction for Question Answering*, Doctoral Dissertation, University of Central Florida Orlando, Florida, 2006.

[27] N. A. Bennett, Q. He, C. T. K. Chang, and B. R. Schatz, *Concept Extraction in the Interspace Prototype*, Urbana Champaign, 1999.

[28] B. Gelfand, M. Wulfekuler, and W. Punch, "Automated concept extraction from plain text," *in Proc. AAAI Workshop on Text Categorization*, pp. 13-17, Jul. 1998.

[29] P. M. Ramirez and C. A. Mattmann, "ACE: improving search engines via automatic concept extraction," in *Proc. Int. Conf. on Information Reuse and Integration, IRI*, pp. 229-234, Nov. 2004.

[30] S. Mohamadi, K. Badie, and A. Moeini, "Using frame-based lexical chains for extracting key points from texts," in *Proc. the 3rd Int. Conf. on Creteave Content Technologies*, *CONTENT'11,* pp. 68-73, Sept. 2011.

[31] R. Barzilay and M. Elhadad, "Using lexical chains for text summarization," in I. Mani and M. T. Maybury, Eds., *Advances in Automatic Text Summarization*, pp. 111-121, The MIT Press, Cambridge, 1999.

[32] S. Mohamadi and K. Badie, "Extracting key concept from English texts by the use of FrameNet," in *Proc. 17th National CSI Computer Conf.*, pp. 384-389, , Mar. 2012. [in Persian]

[33] M. Ajgalik, M. Barla, and M. Bielikova, "From ambiguous words to key-concept extraction," in *Proc.-Int. Workshop on Database and Expert Systems Applications, DEXA'13*, pp. 63-67, Aug. 2013.

[34] Y. Liu, M. Shi, and C. Li, "Domain ontology concept extraction method based on text," in *Proc. 15th In. Conf. on Computer and Information Science, ICIS'16*, 5 pp., Jun. 2016.

[35] M. Hearst, "TextTiling: segmenting text into multi-paragraph subtopic passages," *Computational Linguistics*, vol. 23, no. 1, pp. 33-64, Mar. 1997.

[36] F. Choi, "Advances in domain independent linear text segmentation," in *Proc. of the 1st North American Chapter of the Association for Computational Linguistics Conf.*, pp. 26-33, Apr. 2000.

[37] D. Das, D. Chen, A. Martins, N. Schneider, and N. A. Smith, "Frame-semantic parsing," *MIT Press J., Comput. Linguist.*, vol. 40, no. 1, pp. 9-56, Mar. 2014.

[38] K. Erk and S. Pado, "Shalmaneser-a toolchain for shallow semantic parsing," in *Proc. of LREC*, vol. 6, pp. 527-532, May 2006.

[39] R. Johannson, *Language Technology at LTH*, Lund University. http://nlp.cs.lth.se/software. Accessed on 12 Dec. 2016.

[40] R. Johansson and P. Nugues, "LTH: semantic structure extraction using nonprojective dependency trees," in *Proc. of the 4th Int. Workshop on Semantic Evaluations, Association for Computational Linguistics*, pp. 227-230, Jun. 2007.

[41] D. Das, *No Title*, Noah Smith's NLP Group at Carnegie Mellon University, http://www.ark.cs.cmu.edu/SEMAFOR. Accessed on 2 Dec. 2016.

[42] D. Das, N. Schneider, D. Chen, and N. A. Smith, *SEMAFOR 1.0: A Probabilistic Frame-Semantic Parser*, Lang. Technol. Institute, Sch. Comput. Sci. Carnegie Mellon Univ., 2010.

[43] W. Gale, K. W. Church, and D. Yarowsky, "Estimating upper and lower bounds on the performance of word-sense disambiguation programs," in *Proc. of 30th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics*, pp. 249-256, Jun. 1992.

[44] H. G. Silber and K. F. McCoy, "Efficient text summarization using lexical chains," in *Proc. of 5th Int. Conf. on Intelligent User Interfaces*, pp. 252-255, Jan. 2000.

سودابه محمدی در سال ۱۳۸۶ مدرک کارشناسی مهندسی کامپیوتر (گرایش نرم افزار) خود را از دانشگاه رازی کرمانشاه و در سال ۱۳۸۹ مدرک کارشناسی ارشد خود را در رشته‌ی مهندسی کامپیوتر (گرایش الگوریتم و محاسبات) از دانشگاه تهران اخذ نمود. وی از سال ۱۳۹۱ تاکنون در دانشکده فناوری اطلاعات دانشگاه صنعتی کرمانشاه به عنوان عضو هیأت علمی گروه مهندسی کامپیوتر مشغول به فعالیت است. زمینه‌های تحقیقاتی مورد علاقه نام‌برده متنوع بوده و در زمینه‌های پردازش زبان طبیعی، بازیابی متون، محاسبات سبز و طراحی تجزیه‌گر زبان فارسی می‌باشد.

کامبیز بدیع مدارک کارشناسی، کارشناسی ارشد و دکترای خود را در رشته مهندسی برق، گرایش بازشناسی الگو از انستیتو تکنولوژی توکیو، ژاپن دریافت نمودند. ایشان عضو هیأت علمی پژوهشگاه ارتباطات و فناوری اطلاعات و در حال حاضر معاون پژوهش و توسعه ارتباطات علمی، پژوهشگاه ارتباطات و فناوری اطلاعات می‌باشند. درضمن، ایشان استاد وابسته دانشکده علوم مهندسی، دانشکده فنی، دانشگاه تهران و عضو مدعو شاخه مهندسی برق و کامپیوتر، فرهنگستان علوم هستند. زمینه‌های پژوهشی مورد علاقه نام‌برده عبارتند از: یادگیری ماشین (ML) و مدلسازی شناختی به طور اعم، و مدلسازی رایانشی فرآیندهای قیاسی، فرآیندهای تفسیری و فرآیندهای مبتنی بر تجربه به طور اخص، با تاکید بر مقاصدی از قبیل انتقال بهینه مفاهیم و ایجاد تکنیک‌های نوین، و هوش معنایی با تاکید بر متن کاوی و ایجاد ایده‌های جدید.