علوم و مهندسی کامپیوتر
نشریه علمی پژوهشی انجمن کامپیوتر ایران
مجلد ۶، شماره ۳ (الف)، ۱۳۸۷
صفحات ۴۰-۵۲
مقاله عادی

اجتماع نمونه‌دهی هستان‌شناسی و حاشیه‌نویسی معنایی متون فارسی در سیستم POPTA

مهرنوش شمس‌فرد بهاره صراف‌زاده

دانشکده مهندسی برق و کامپیوتر، دانشگاه شهید بهشتی، تهران، ایران

چکیده

نمونه‌دهی خودکار هستان‌شناسی به استخراج نمونه‌های کلاس‌های یک هستان‌شناسی از متن و افزودن آنها به هستان‌شناسی می‌پردازد. از سوی دیگر وب معنایی با افزودن حاشیه‌نویسی به اسناد وب کنونی محتوای ساخت‌یافته‌ای را فراهم می‌نماید. حاشیه‌نویسی‌های معنایی داده‌های نمونهٔ کلاس‌های هستان‌شناسی را برچسب‌گذاری نموده و آنها را بر کلاس‌های هستان‌شناسی منطبق می‌نمایند. بنابراین فعّالیت نمونه‌دهی به هستان‌شناسی می‌تواند با ایجاد حاشیه‌نویسی‌های معنایی همراه باشد.

در این مقاله یک سیستم نمونه‌دهی و حاشیه‌نویسی میان زبانی[۱] به نام POPTA[۲] معرفی می‌شود که متون فارسی را با توجه به یک هستان‌شناسی (با مداخل لغوی انگلیسی) حاشیه‌نویسی کرده و این هستان‌شناسی را نیز با توجه به این متون نمونه‌دهی می‌نماید. این سیستم از ترکیبی از روش‌های آماری و مبتنی بر الگو به همراه متدهای مبتنی بر وب، موتورهای جستجو و دایرةالمعارف آنلاین ویکی‌پدیا به منظور نمونه‌دهی به هستان‌شناسی و حاشیه‌نویسی متون بهره می‌گیرد.

کلمات کلیدی: نمونه‌دهی هستان‌شناسی، حاشیه‌نویسی متون، پردازش زبان فارسی، یادگیری هستان‌شناسی، ویکی‌پدیا، وب معنایی.

۱- مقدمه

بعدی وب – که وب معنایی نامیده می شود – برای بهبود معنای داده‌ها از هستان‌شناسی‌ها بهره می گیرد. ساخت دستی این هستان‌شناسی‌ها بسیار زمان بر و پر هزینه است. به علاوه هستان‌شناسی‌های ساخته شده توسط انسان‌ها به شیوه تفکر و سلایق فردی آنها وابسته است. از این رو به دنبال بهره گیری از روش‌های خود کار یادگیری هستان‌شناسی (OL) و استخراج دانش می‌باشیم . یکی از جنبه های ساخت خودکار هستان‌شناسی‌ها تکمیل هستان‌شناسی با نمونه‌های مفاهیم موجود در آن است. این فعالیت نمونه‌دهی هستان‌شناسی[۳] (OP) نامیده می شود.

از سوی دیگر تحقّق یافتن وب معنایی به در دسترس بودن همه جانبهٔ حاشیه‌نویسی معنایی برای اسناد موجود و جدید برروی وب نیازمند است. حاشیه‌نویسی‌های معنایی داده‌های نمونهٔ کلاس‌های هستان‌شناسی را برچسب‌گذاری نموده و آنها را بر کلاس‌های هستان‌شناسی منطبق می‌نمایند.

بنابراین فعّالیت نمونه‌دهی به هستان‌شناسی با ایجاد حاشیه‌نویسی های معنایی همراه است. ساخت کاملاً خودکار حاشیه‌نویسی‌های معنایی مسئله‌ای حل نشده است. در عوض سیستم‌های امروزی بر ساخت نیمه‌خودکار حاشیه‌نویسی‌ها متمرکز شده‌اند [۱].

به عبارت دیگر با داشتن یک هستان‌شناسی دامنه (مفاهیم و روابط) و استخراج نمونه‌های این مفاهیم یا کلاس‌ها از یک متن، می‌توان این نمونه‌ها را در متن برچسب‌گذاری معنایی کرد (حاشیه‌نویسی) و یا این نمونه ها را تحت کلاس پدر در هستان‌شناسی متصل نمود (نمونه‌دهی).

تا کنون رویکردهایی برای حاشیه‌نویسی و یا نمونه‌دهی ارائه شده است. بطور کلّی حاشیه‌نویسی معنایی، به یک هستان‌شناسی دارای دامنه و زبان یکسان با متن نیاز دارد. با وجود این که هستان‌شناسی‌ها پایگاه‌های دانش مفهومی هستند و بطور نظری باید مستقل از زبان باشند، هنگامی که از نمادها برای بازنمایی مفاهیم استفاده می‌کنیم، معمولاً از واژگان[۴] یک زبان برای نشان دادن مفاهیم بهره می‌بریم و بنابراین ویژگی استقلال از زبان را از دست می‌دهیم. به بیان دیگر عناصر

هستان‌شناسی با یک زبان خاص (معمولاً انگلیسی) واژه‌سازی[۵] می‌شوند و نمی‌توانند مستقیماً برای پردازش متون فارسی بکار روند.

بنابراین از سویی برای حاشیه‌نویسی متون فارسی نیاز به هستان‌شناسی های واژه سازی شده با واژه های فارسی (یا هستان‌شناسی‌های فارسی) داریم که درحال حاضر موجود نیستند و لازم است بطور دستی یا خودکار تولید شوند. از سوی دیگر یکی از راه‌های یادگیری و نمونه‌دهی هستان‌شناسی‌های فارسی استفاده از پیکره‌های برچسب خورده معنایی است که متاسفانه برای زبان فارسی تهیه نشده اند. این مسئله منجر به ایجاد حلقه‌ای شده است که برای رفع آن، در این مقاله به ارائه یک رویکرد موازی میان زبانی[۶] می‌پردازیم.

در این رویکرد از وب نیز به عنوان بزرگترین و گسترده‌ترین پیکرهٔ موجود، از گوگل[۷] به عنوان قدرتمندترین موتور جستجوی کنونی و از ویکی‌پدیا[۸] به عنوان یک پایگاه اطلاعاتی جامع که به تازگی صفحات فارسی آن نیز در حال گسترش است به منظور افزایش کارایی و دقت سیستم و نیز غلبه بر گلوگاه اکتساب دانش بهره گرفتیم.

POPTA نه تنها نخستین سیستم حاشیه‌نویسی متو.ن فارسی و نمونه‌دهی هستان‌شناسی از روی متون فارسی است، بلکه جزء معدود سیستم‌هایی در جهان است که دو فرآیند نمونه‌دهی و حاشیه‌نویسی را به صورت میان زبانی و بطور تواماً انجام می‌دهد.

علاوه بر ویژگی‌های گفته شده، سیستم POPTA خود را بر همه فرمت‌های کاراکترهای موجود در کلمات فارسی منطبق نموده و برای کار با هر منبع، فرمت کلمات ورودی را با فرمت کلمات موجود در آن منبع تطبیق می‌دهد.

در ادامه این مقاله در بخش ۲ مروری بر کارهای صورت گرفته برای نمونه‌دهی و حاشیه‌نویسی هستان‌شناسی‌ها خواهیم داشت، در بخش ۳ سیستم POPTA را به عنوان یک سیستم نمونه‌دهی و حاشیه‌نویسی فارسی معرفی نموده و در بخش ۴ به ارزیابی آن می‌پردازیم. در آخر نیز نگاهی به رویکردهای آینده خواهیم داشت.

۲- کارهای مرتبط

در این بخش به معرفی فعالیت‌های انجام شده در حوزه نمونه‌دهی و حاشیه‌نویسی معنایی خواهیم پرداخت. با مطالعه فعالیت‌های انجام شده در حوزه نمونه‌دهی هستان‌شناسی‌ها چهار معیار نوع منبع ورودی، نمونه‌های استخراج شده، رویکرد نمونه‌دهی و محصول نهایی را جهت دسته بندی روش های مختلف استخراج نموده‌ایم [۲]. در ادامه این بخش ابتدا به دسته بندی روشهای موجود با توجه به این معیارها پرداخته و در هر دسته مثال هایی از سیستم های نمونه‌دهی ارائه خواهیم نمود.

معیار اول: منبع ورودی – منظور از منبع ورودی منبعی است که نمونه‌ها از آن استخراج می شوند. این منبع می‌تواند یک پیکره متنی (مانند منبع ورودی در سیستم KMi [۳]) و یا صفحات موجود در وب (مانند منبع ورودی در [۴، ۵].) باشد. در میان سیستم هایی که عمل نمونه‌دهی و حاشیه‌نویسی را به کمک وب انجام می دهند، گروهی از سیستم‌ها از سرویس‌های موجود در سطح وب – مانند موتور جستجوی گوگل– بهره گرفته و گروهی دیگر از منابع اطلاعاتی موجود در سطح وب – مانند دایرةالمعارف آنلاین ویکی‌پدیا – برای استخراج دانش مورد نیاز خود استفاده کرده‌اند. OntoGenie [۴] یک ابزار نیمه اتوماتیک است که هستان‌شناسی‌های خاص قلمرو و داده‌های غیر ساخیافته در وب را به عنوان ورودی دریافت کرده و نمونه‌های هستان‌شناسی را تولید می‌نماید. این ابزار از هستان‌شناسی زبانی وردنت به عنوان پلی بین هستان‌شناسی‌های خاص قلمرو و داده‌های وب بهره می‌گیرد. PANKOW [۶] سیستم حاشیه‌نویسی و نمونه‌دهی دیگری است که معتقد است الگوهای ساختواژی مشخص که در متن پیدا می‌شوند، حامل یک رابطهٔ معنایی خاص هستند. این سیستم از الگوهای هیرست[۹] [۷] استفاده می‌کند سیستم معرفی شده در [۸] با استفاده از الگوهای رابطهٔ خاص قلمرو دست‌ساز – که تعمیم یافتهٔ الگوهای هیرست هستند – به نمونه‌دهی یک هستان‌شناسی می‌پردازد. این الگوریتم از نمونه‌های یک کلاس که توسط گوگل برگردانده شده‌اند، برای یافتن نمونه‌های کلاس‌های دیگر بهره می‌گیرد.

بانسکو و پاسکا [۹] از ویکی‌پدیا به منظور ایجاد یک روش رفع ابهام موجودیت‌های نامدار استفاده کرده و با استفاده از یک دیکشنری، اسامی خاص را به موجودیت‌های نامدار مربوط به آنها متصل می‌نمایند. سیستم آنها برای استخراج موجودیت‌های نام‌دار از صفحه‌های تغییر مسیر و رفع ابهام و نیز سیستم رده‌ها و پیوندهای ویکی‌پدیا بهره می‌گیرد. چرنو و همکارانش [۱۰] از این ایده استفاده می‌کنند که طبقات یا رده‌هایی که ارتباطات و پیوندهای زیادی با هم دارند، دارای روابط معنایی با هم هستند.

معیار دوم: نمونه‌های استخراج شده آنچه تحت عنوان نمونه از منبع ورودی استخراج و به مفاهیم متناظر در هستان‌شناسی متصل می شود، ممکن است در سیستم های مختلف متفاوت باشد. دو گونه متداول از این نمونه ها صفحات وب و موجودیت های نامدار متون زبان طبیعی هستند. در گونهٔ اول در یک قلمرو مشخص، صفحات وب نمونه‌هایی از هستان‌شناسی آن دامنه هستند. از این رو ما نیازمند ابزارهایی نیمه‌خودکار برای یافتن این نمونه‌ها و روابط میان آنها و مقداردهی به ویژگی‌های آنها می‌باشیم. برای مثال سیستم OILSW [۱۱] از صفحات وب به عنوان نمونه استفاده می‌نماید و هر صفحه را به مفهوم معادلش در یک هستان‌شناسی پیش فرض متصّل می‌نماید.

در گونهٔ دوم نمونه‌ها، عناصر متنی (مانند کلمات و عبارات) هستند که متناظر با نمونه های جهان خارجند و باید به مفهوم معادل خود در هستان‌شناسی مربوطه متّصل شوند. برای مثال در سیستم KMi رویدادها به عنوان کلاس‌های هستان‌شناسی تعریف شده‌اند. در این سیستم ابتدا توپولوژی رویداد مستقیماً در هستان‌شناسی تعریف می‌شود. سپس برای هر رویداد، اسلات‌هایی تعریف می‌گردد که ممکن است با یک مولفهٔ استخراج اطلاعات نمونه‌دهی شوند. هدف پرکردن خودکار بیشترین تعداد ممکن از حفره‌هاست.

معیار سوم: رویکرد نمونه‌دهی – رویکردهای نمونه‌دهی به دسته‌های اصلی مبتنی بر الگو، آماری و مبتنی بر مکاشفه تقسیم می‌شوند. در رویکردهای مبتنی بر الگو، نمونه‌دهی هستان‌شناسی با استفاده از الگوها و یا برپایهٔ ساختار ترم‌ها انجام می‌شود. این رویکردها بدنبال عباراتی هستند که بوضوح وجود یک رابطهٔ is-a را بین دو کلمه نشان می‌دهند، مانند "The ant is an insect" یا "ants and other insects". الگوهای هیرست [۷] نمونه‌های معروفی برای استخراج روابط is-a از متن هستند که در بسیاری سیستم ها [۶، ۱۲] مورد استفاده قرار گرفته‌اند.

در رویکردهای مبتنی بر مکاشفه که در بسیاری موارد به صورت ترکیبی با رویکردهای دیگر مورد استفاده قرار می‌گیرند، قوانین مکاشفه‌ای برای استخراج روابط شمول بکار می‌روند. برای مثال [۱۰] از یک مکاشفه انطباق هسته[۱۰] استفاده می‌کند که بر اساس آن میان هسته یک عبارت و کل عبارت می‌توان یک رابطهٔ is-a برقرار نمود؛ مثلاً "Christmas tree" یک نوع از "tree" است.

در روش‌های آماری از ویژگی‌های وابسته به متن استفاده می‌شود [۱۴]. این رویکردها از یک پیکره برای استخراج ویژگی‌هایی از متن که یک کلاس معنایی را آشکار می‌سازد، استفاده می‌کنند. این ویژگی‌ها می‌توانند ظاهری یا ساختاری باشند. ارزیابی انجام‌شده برای مقایسه نشان داده است که ویژگی‌های ساختاری به کارایی بهتری منجر می‌شوند. همچنین در بسیاری موارد سیستم‌های نمونه‌دهی از ترکیبی از رویکردهای فوق بهره می‌گیرند. سیستم معرفی شده در [۱۲] از الگوهای هیرست، سلسله مراتب وردنت[۱۱]، مکاشفه‌های انطباق هسته، استنتاج

مبتنی بر پیکره و دیگر منابع موجود برای یادگیری روابط طبقه‌ای استفاده می‌نماید.

همچنین رویکردهای اخیر را می‌توان با توجّه به استفادهٔ متفاوت از داده‌های آموزشی، به دو گروه تقسیم نمود: رویکردهای بدون نظارت و رویکردهای با نظارت که از داده‌های آموزشی برچسب‌گذاری شدهٔ دستی استفاده می‌کنند. در حالی که روش‌های بدون نظارت کارایی پایینی دارند، رویکردهای بانظارت به دقّت بالاتری دست یافته‌اند، ولی به ساخت دستی یک مجموعهٔ آموزشی که در عین حال آنها را از کاربردهای در مقیاس بزرگ باز می‌دارد، نیاز دارند [۱۴].

معیار چهارم: محصول نهایی- سیستم‌های موجود را بسته به نوع خروجی‌ای که تولید می‌نمایند، می‌توان به دو گروه سیستم‌های نمونه‌دهی و سیستم‌های پشتیبان تقسیم نمود. گروه اول سیستم‌هایی هستند که برای نمونه‌دهی به هستان‌شناسی‌ها کاربرد دارند و خروجی آنها نمونه‌های تولید شده برای مفاهیم هستان‌شناسی ورودی است (مانند [۴، ۸، ۱۱، ۱۴، ۱۵]). گروه دوم نیز ملزومات کار گروه اول را فراهم می‌نمایند. درواقع این سیستم‌ها به نمونه‌دهی هستان‌شناسی‌ها نمی‌پردازند ولی پیش‌نیازهای لازم برای فعالیت‌های سیستم‌های نمونه‌دهی را فراهم می‌نمایند. این سیستم‌ها عموماً به کار استخراج موجودیت‌های نام‌دار و یا زیرمجموعه‌ای خاص از نمونه‌ها می‌پردازند (مانند [۱۶]).

برخی از سیستم‌های نمونه‌دهی [۶] در واقع سیستم‌های حاشیه‌نویسی هستند که عمل نمونه‌دهی را نیز انجام می‌دهند. در مقابل سکوهای حاشیه‌نویسی معنایی (SAP[۱۲]) وظیفه اصلی یافتن موجودیت‌های نامدار متن و الصاق برچسب معنایی مناسب (با توجه به یک هستان‌شناسی موجود) به آنها را دارا می‌باشند. این سکّوها در معماری، ابزارها و روش‌های استخراج اطلاعات، هستان‌شناسی اولّیه، میزان کار دستی لازم برای انجام حاشیه‌نویسی، کارایی و دیگر ویژگی‌ها مانند مدیریت حافظه با هم متفاوت هستند. این سکّوها می‌توانند بر مبنای نوع روش حاشیه‌نویسی مورد استفاده به دو گروه اصلی مبتنی بر الگو و مبتنی بر یادگیری ماشینی و یا ترکیبی از هردو دسته‌بندی شوند[۱]. SAPهای مبتنی بر الگو ممکن است قادر به کشف الگوهای جدید و یا استفاده از الگوهای از پیش تعریف شده باشند. بیشتر روش‌های کشف الگو یک مجموعهٔ اولّیه از موجودیت‌ها را تعریف و پیکره را به منظور یافتن الگوهایی که این موجودیت‌ها در آنها وجود دارند، پیمایش می‌کنند. موجودیت‌های جدید همراه الگوهای جدید کشف می‌شوند. این فرآیند بصورت بازگشتی تا جایی ادامه می‌یابد که دیگر هیچ موجودیتی کشف نشود یا این‌که کاربر فرآیند را متوقف نماید. حاشیه‌نویسی‌ها می‌توانند توسط قوانینی دستی برای یافتن موجودیت‌ها در متن نیز تولید شوند.

SAPهای مبتنی بر یادگیری ماشینی از دو رویکرد بهره می‌گیرند: احتمال و استنتاج. SAPهای احتمالاتی از مدل‌های آماری برای پیش‌بینی مکان موجودیت‌ها در داخل متن استفاده می‌کنند. برای مثال در الگوریتم DATAMOLD [۱۷] برای یافتن نمونه‌های داده‌ای در داخل صفحات HTML، از مدل مخفی مارکوف استفاده می‌شود. همچنین الگوریتم LP^2 [۱۸] هستهٔ الگوریتم استخراج اطلاعات (IE) در ابزار Amilcare [۱۹] است که برای اعمال استنتاج پوششی[۱۳] توسط SAPهای Armadillo [۲۰] و Ont-O-Mat [۱۹] استفاده می‌شود [۱].

سکّوی KIM [۲۱] نمونه ای مطرح از سکوهای حاشیه‌نویسی است که شامل یک هستان‌شناسی (KIMO)، یک پایگاه دانش، یک حاشیه‌نویس معنایی، یک سرور شاخص گذاری و بازیابی به همراه پایانه‌هایی برای واسطه‌گری با سرور می‌باشد. KIMO یک مجموعه پایه از کلاسهای موجودیت، روابط و محدودیت ویژگیها را تعریف می‌کند. پایگاه دانش به کمک ۸۰۰۰۰ موجودیت، شامل موقعیت‌ها و سازمان‌ها که از یک بدنه خبری عمومی جمع آوری شده‌اند، گسترش می‌یابد. موجودیت‌های دارای نامی که در طی روند حاشیه‌نویسی پیدا شده‌اند، با نوعشان در آنتولوژی و همچنین با یک مرجع در پایگاه دانش تطبیق داده می‌شوند. مولفهٔ استخراج اطلاعات حاشیه‌نویسی معنایی با استفاده از مولّفه‌های ابزار GATE اعمال می‌شود.

سیستم دیگر، کار ارائه شده توسط کاسادو و همکارانش[۲۲، ۲۳] است. در این کار روشی برای حاشیه‌نویسی خودکار روابط معنایی مختلف در ویکی‌پدیا ارائه شده است. این فرآیند مبتنی بر کشف و تعمیم خودکار الگوهای واژگانی بوده و امکان تشخیص روابط موجود میان مفاهیم را می‌دهد. بدین منظور از یک رویکرد مبتنی بر یادگیری خودکار الگوهای واژی- نحوی[۱۴] بهره گرفته شده. این فرآیند با فهرست اولیه‌ای که شامل زوج‌هایی از آیتم‌های مرتبط است، آغاز می‌گردد. سپس جملات بسیاری که دربرگیرندهٔ این زوجها هستند بطور خودکار از وب جمع‌آوری شده و توسط ابزارهای NLP مانند قطعه‌بند[۱۵]، برچسب‌زن مقولهٔ نحوی، ریشه‌یاب و تشخیص‌دهنده موجودیت‌های نامدار پردازش می‌شوند. اطلاعاتی که از این پردازش بدست می‌آیند، می‌توانند برای تحقیق در مورد لغات، ساختارها و موجودیت‌هایی که عموماً هنگامی که رابطه‌ای بین دو مفهوم بیان می‌شود در زبان طبیعی استفاده می‌شود، بکار روند. این سیستم بر روی ۸ رابطهٔ سال تولّد شخص، سال مرگ شخص، مکان تولّد شخص، بازیگر – فیلم، نویسنده – کتاب، بازیکن فوتبال – تیم، کشور – رئیس دولت، کشور – پایتخت تست گردیده است.

همچنین در سال‌های اخیر سیستم‌های مختلفی برای حاشیه‌نویسی متون خاص قلمرو جهت گسترش وب معنایی بوجود آمده‌اند که از جمله آنها می‌توان به [۲۴] برای حاشیه‌نویسی متون حوزه زیست پزشکی اشاره نمود.

۳- معرفی سیستم POPTA

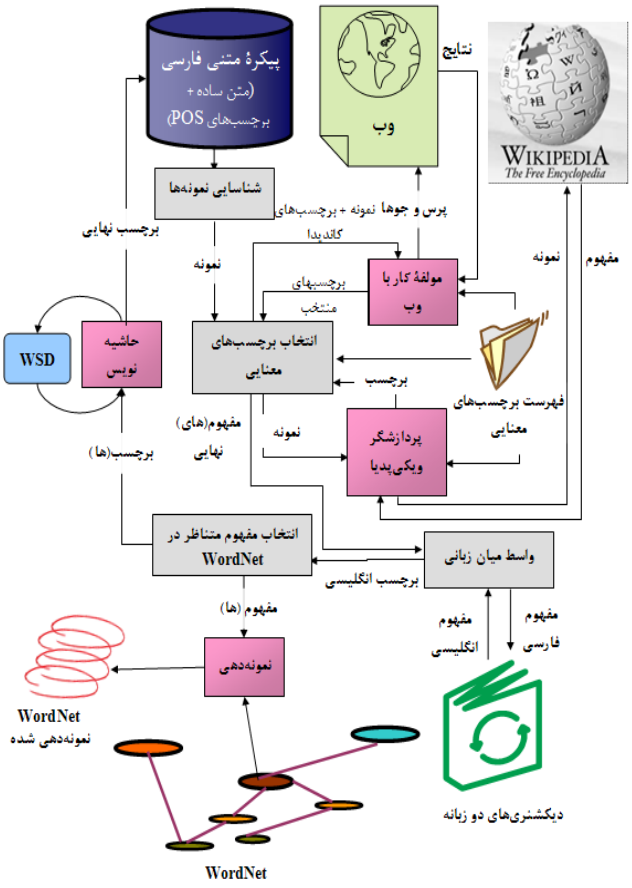
POPTA یک سیستم موازی نمونه‌دهی و حاشیه‌نویسی است که بر روی منابع زبانی چندگانه کار می‌کند. ورودی‌های این سیستم یک هستان‌شناسی موجود با واژه سازی انگلیسی (مانند SUMO یا وردنت)، یک پیکرهٔ متنی از متون زبان فارسی که برچسب مقولهٔ نحوی دارند و نیز صفحات وب به زبان فارسی خواهد بود. هدف این سیستم یافتن نمونه‌های مناسب از پیکره و نیز از روی وب برای نمونه‌دهی به هستان‌شناسی ورودی و در نهایت حاشیه‌نویسی معنایی متون فارسی می‌باشد.

الگوریتم کلی کار به این ترتیب است: ابتدا کلیهٔ اسامی خاص موجود در پیکرهٔ متنی فارسی به عنوان نمونه استخراج شده و با توجه به روش‌هایی که در بخش‌های بعدی به آنها خواهیم پرداخت، مجموعه‌ای از برچسب‌های اولیه به آنها تخصیص می‌یابد. سپس این نمونه‌ها به همراه مجموعهٔ برچسب‌های کاندیدای آنها وارد مولفهٔ کار با وب (Googling) شده و با استفاده از موتور جستجوی گوگل (در صورت امکان) مناسبترین برچسب ممکن به هر یک از این نمونه‌ها تخصیص می‌یابد. در صورت موفقیت‌آمیز بودن عملیات انتخاب مناسب‌ترین برچسب، این برچسب‌ها توسط یک واسط میان‌زبانی – که از دو دیکشنری دو زبانهٔ فارسی به انگلیسی و انگلیسی به فارسی استفاده می‌کند – به زبان انگلیسی ترجمه شده و وارد فاز انتخاب مفهوم متناظر در هستان‌شناسی مقصد (وردنت) می‌شوند. پس از یافتن مفهوم متناظر با هر برچسب، از این مفاهیم برای حاشیه‌نویسی متن ورودی استفاده می‌شود. در صورت عدم موفقیت در انتخاب مناسب‌ترین برچسب به کمک گوگل، نمونهٔ ورودی به عنوان یک پرس و جو به ویکی‌پدیا داده می شود تا نوع آن تعیین گردد و ادامهٔ کار از مولفهٔ یافتن مفهوم متناظر در وردنت و مانند بخش قبل دنبال می‌گردد.

لازم به ذکر است که از آنجایی که کاراکترهای فارسی دارای کدگذاری‌های مختلفی می‌باشند و در برخی نوشته ها از کدهای فارسی و دربرخی از کدهای عربی استفاده می شود، در متون فارسی شاهد برخی حروف مشابه با کدهای متفاوت هستیم. این مشکل بیشتر برای حروف ‘ی’ و ‘ک’ مشهود است. در برخی

از کلمات این حروف به شکل نرمال 'ی' و 'ک' و در برخی دیگر بصورت 'ي' و 'ك' نوشته شده‌اند. بنابراین در هرکجای سیستم که به نوعی با متون فارسی موجود سر و کار داریم (مانند متون پیکره، اطلاعات فارسی موجود در وب و یا دیکشنری دوزبانه) لازم است با انجام عملیات نرمال‌سازی فرمت‌های غیرهمسان را همسان نماییم و یا کلیه فرمت‌های ممکن را جهت انطباق بیازماییم. این کار سرعت پردازش را پایین می‌آورد ولی در مقابل درصد خطا را نیز کاهش می‌دهد.

معماری این سیستم در شکل ۱ نمایش داده شده است. در ادامه به شرح عملکرد هر یک از این مولفه‌ها در سیستم POPTA می‌پردازیم.



شکل ۱- معماری سیستم POPTA

۳-۱- پیکره

این سیستم از یک پیکره متنی فارسی برچسب خورده (پیکره دکتر بیجن‌خان) استفاده می‌کند. در این پیکره برای هر کلمه اطلاعاتی از قبیل مقولهٔ نحوی، مفرد یا جمع بودن، کسرهٔ اضافه، ضمایر متصل، اسامی خاص یا عام و ... فراهم است. POPTA از برچسب‌های مقولهٔ نحوی[۱۶]، مفرد یا جمع بودن و خاص یا عام بودن اسامی استفاده می‌نماید.

۳-۲- شناسایی نمونه‌ها

بسیاری از سیستم‌های نمونه‌دهی به هستان‌شناسی‌ها در گام اول از یک سیستم IE برای استخراج موجودیت‌های نامدار موجود در متن بهره می‌گیرند. از آنجایی که در حال حاضر سیستم استخراج موجودیت‌های نامدار برای متون زبان فارسی وجود ندارد، لازم است وظایف چنین سیستمی در داخل سیستم نمونه‌دهی به انجام رسد. بطور کلّی برای یافتن کاندیداهای مناسب برای نمونه‌ها و یا مفاهیم جدید – که باید حاشیه‌نویسی شوند – در یک صفحهٔ وب، ابتدا لازم است محتوای متنی صفحات وب را استخراج کرده (برچسب‌های HTML را حذف کنیم) و یک برچسب زن مقولهٔ نحوی بر روی آن اعمال کنیم تا به هر قطعه[۱۷] ردهٔ دستوری متناظر با آن، تخصیص یابد. این برچسب‌زن می‌تواند اسامی خاص را تعیین نماید.

در متون لاتین معمولاً می‌توان نمونه‌ها را از طریق کلماتی که با حروف بزرگ نوشته شده‌اند و برچسب اسم خاص[۱۸] دارند، جستجو کرد. برای یافتن مفاهیم نیز می‌توان از کلمات با حروف کوچک و با برچسب اسم عام[۱۹] بهره برد. در زبان‌هایی مانند فارسی، عربی و ... که فاقد حروف بزرگ هستند باید اطلاعات ساخت - واژی[۲۰] و بویژه اطلاعات مقولهٔ نحوی در مورد هر کلمه فراهم باشد. در POPTA برای استخراج نمونه‌ها، از متون یک پیکرهٔ فارسی حاوی برچسب‌های مقولهٔ نحوی به عنوان ورودی استفاده شده است. این سیستم کلمات مفرد دارای برچسب اسم خاص (PR) را به عنوان نمونه استخراج می‌نماید. انواع دیگری از اسامی خاص وجود دارند که در این پیکره با برچسب اسم خاص مشخص نشده‌اند. از جملهٔ این اسامی می‌توان به نام کتاب‌ها، مجلّات، سریال‌ها، فیلم‌ها و سایر محصولات فرهنگی – هنری اشاره نمود. این سیستم برای بازیابی این اسامی از برخی علائم نگارشی خاص مانند قرار گرفتن نام در "" و یا عباراتی مانند "کلمهٔ ۱: کلمهٔ ۲" که می‌تواند نشان‌دهندهٔ رابطهٔ "کلمهٔ ۲ نوعی (is-a) کلمهٔ ۱ است" باشد، استفاده می‌نماید. همچنین وجود برخی کلمات کلیدی در همسایگی این اسامی می‌توانند در انتخاب آنها به عنوان اسامی خاص مؤثر باشند. در گام‌های بعدی می‌توان محدودهٔ نمونه‌ها را گسترش داد و انواع اسامی از جمله اسامی عام و نیز برخی از افعال را نیز به عنوان نمونه در نظر گرفت. بدیهی است با توسعهٔ سیستم‌های استخراج اطلاعات برای متون فارسی می‌توان از این سیستم‌ها برای استخراج خودکار نمونه‌ها از سایر متون فارسی و نیز متون موجود در سطح وب – که فاقد برچسب‌های مقولهٔ نحوی هستند – بهره برد.

۳-۳- انتخاب برچسب‌های معنایی[۲۱]

در سیستم POPTA، به منظور برچسب‌گذاری معنایی متون، مجموعه‌ای از برچسب‌های معنایی تهیه شده است. از جملهٔ این برچسب‌ها می‌توان به انواع مکان‌های جغرافیایی، انسان‌ها، انواع محصولات فرهنگی – هنری، دوره‌های تاریخی (سلسله‌ها، دودمان‌ها، خاندان‌ها)، فرقه‌ها، اقوام، مذاهب و ادیان، زبان‌ها، ماه‌های سال، سیّارات و ... اشاره نمود.

همانطور که مشاهده می‌شود تعداد و تنوع این برچسب‌های معنایی بسیار زیاد است (در نسخه فعلی سیستم ۵۴ برچسب از پیش تعریف شده دارد) و برای تعیین برچسب مناسب برای هر نمونه باید به نحوی تعداد برچسب‌های کاندیدا برای آن نمونه را بوسیله یک الگوریتم برچسب‌زنی کاهش داد تا به حد ایده‌آل یک برچسب برای هر نمونه نزدیک شویم.

این سیستم از مجموعه‌ای از قوانین مکاشفه‌ای (مانند قانون انطباق هسته[۲۲] در گروه‌های اسمی) به همراه ترکیبی از روش‌های آماری از جمله روش‌های N-gram و روش‌های مبتنی بر الگو (براساس الگوهای ارائه شده در [۲۵]) برای محدود کردن لیست مفاهیم مرتبط با هر نمونه بهره می‌گیرد.

در بخشی از این رویکرد ترکیبی، سیستم، هم وقوعی کلمات کلیدی نشان‌دهنده یکی از کلاس‌های هستان‌شناسی با نمونه کاندیدای موردنظر را بررسی می‌کند. برای یافتن این کلمات کلیدی می‌توان به سه طریق عمل کرد: (۱) ساخت خودکار فهرستی از ترم‌هایی که در وب هم‌وقوعی بالایی با این نمونه‌های کاندیدا دارند، (۲) انتخاب مفاهیم موجود در سطوح پایین‌تر مفاهیم عمومی که در

هستان‌شناسی‌های خاص آن حوزه و یا در هستان‌شناسی وردنت موجودند و (۳) استفاده از فهرست‌های موجود در ویکی پدیا.

راه‌حل اول که درواقع یک روش آماری است، نیازمند حذف کلمات زائد[۲۳] از این فهرست و اولویت دادن به کلماتی می‌باشد که در مفاهیم یک هستان‌شناسی وجود دارند و نشان‌دهندهٔ کلمات عمومی‌تر هستند. راه دوم با مفاهیم موجود در ساختار هستان‌شناسی‌ها کار می‌کند و نیازمند پیمایش گراف هستان‌شناسی و یافتن فرزندان مناسبتر به عنوان نماینده‌هایی برای کلاس‌های در بر گیرندهٔ نمونه‌های کاندیداست.

راه حل سوم راه حل بهتری به نظر می‌رسد و پیچیدگی‌های راه‌حل‌های قبلی را ندارد. تنها مشکل این راه این است که صفحات فارسی ویکی‌پدیا هنوز به گستردگی صفحات دیگر زبان‌ها و بویژه صفحات انگلیسی نیستند و این امر باعث می‌شود که فهرست کاملی از تمامی کلمات کلیدی مورد نیاز برای ایجاد برچسب‌های معنایی را بدست ندهد.

از آنجایی که مجموعهٔ کلمات کلیدی مورد نیاز این سیستم برای تشخیص انواع اسامی خاص محدود و از پیش تعریف شده می‌باشد، در حال حاضر سیستم POPTA از فهرست کلمات کلیدی – که بصورت دستی ایجاد شده است – استفاده می‌کند. در روش پیشنهادی سیستم درمیان کلماتی که در پنجره‌ای به طول ۴ و قبل از نمونهٔ کاندیدای موردنظر رخ داده‌اند به جستجوی کلمات کلیدی فوق می پردازد و در صورت یافتن، آنها را به عنوان مفاهیم مرتبط کاندیدا انتخاب می‌نماید.

بطور کلّی کارایی روش‌های N-gram بستگی زیادی به سایز پنجره و ویژگی‌های زبان طبیعی مورد استفاده دارند، بنابراین برای برخی نمونه‌ها (مانند اسامی کشورها، شهرها و سایر مکان‌های جغرافیایی) سایز پنجره را به اندازهٔ طول یک گروه اسمی در نظر گرفته و از مکاشفه‌های انطباق هسته برای کار با هستهٔ گروه‌های اسمی بهره گرفته‌ایم.

روش‌های مبتنی بر الگو از کارایی بیشتری نسبت به روش‌های آماری برخوردارند و درصورت تطابق بخشی از متن با الگوهای مشخص می‌توان روابط معنایی مفیدی را استخراج نمود. از این رو برای بهره‌مندی از این روش‌ها الگوهای مختلفی که نشان‌دهندهٔ روابط is-a و یا instance-of می‌باشند را فرموله کردیم.

به عبارت دیگر این سیستم از سهولت و عمومیت روش‌های آماری همراه با دقت بالا و کارایی بیشتر روش‌های مبتنی بر الگو بهره گرفته است و بر خطاهای ناشی از روش‌های آماری و رخداد پایین الگوهای معنایی در متون پیکره و نیز در سطح وب غلبه می‌نماید.

در مرحلهٔ تعیین برچسب‌های معنایی برای هر نمونه اغلب با بیش از یک برچسب کاندیدا برای نمونه‌ها روبرو هستیم. در واقع روش‌های آماری و الگویی به همراه مکاشفه‌های بکار رفته به تنهایی نمی‌توانند مناسبترین برچسب را برای یک نمونه تعیین نمایند. از این رو در این مرحله یک سری برچسب اولیه به نمونه‌ها تخصیص داده می‌شود که نشان‌دهندهٔ تمامی برچسب‌های کاندیدای مجاز برای آن نمونه می‌باشند. در این مجموعه می توان احتمال رخداد برچسب‌ها را براساس شواهد و کلمات کلیدی که در همسایگی نمونه رخ داده اند تقویت یا تضعیف نمود.

برای مثال دو جمله "امروز، قدرت حکومت در دست سلطان زنگبار است." و "با فتوحات سلطان محمود غزنوی، این منطقه به دست غزنویان افتاد." را در نظر بگیرید. در اولی اسم خاص "زنگبار" که در همسایگی کلمهٔ کلیدی سلطان آمده است، نام یک مکان جغرافیایی است، حال آن‌که در دومی اسم خاص "محمود غزنوی" که باز هم در همسایگی کلمهٔ کلیدی سلطان قرار گرفته است به نام یک انسان اشاره دارد. بنابراین به این دسته از کلمات برچسب اولیهٔ HLOC تخصیص می‌یابد که نشان‌دهندهٔ یکی از دو نوع انسان (H) و یا مکان (LOC) می‌باشد.

نمونهٔ دیگر قرار گرفتن چند اسم خاص در کنار یکدیگر است که می‌تواند در محدوده مورد بررسی نشاندهندهٔ یکی از سه وضعیت زیر باشد:

- همهٔ این اسامی انسان هستند؛ مانند: محمّد حسین شهریار
- همهٔ این اسامی مکان هستند؛ مانند: ملبورن استرالیا
- اسم اول اثری متعلّق به اسم دوم است؛ مانند: شاهنامهٔ فردوسی

در چنین مواردی وجود کلمات کلیدی در همسایگی این عبارات می‌تواند به تعیین قطعی و یا بالا بردن احتمال وجود یک برچسب خاص کمک نماید. برای مثال در موارد فوق (وجود اسامی خاص متوالی) در صورت وجود القابی مانند آقا، خانم، امام، دکتر، مهندس، خاله، سیّد، استاد، جناب، سرکار و ... می‌توان فهمید که این اسامی خاص همگی انسان هستند و بنابراین برچسب Human به آنها تخصیص می‌یابد. در غیر اینصورت اگر کلماتی کلیدی مانند کتاب، نشریه، سریال، مجله، نمایش، تئاتر، فیلم و ... در همسایگی و قبل از اسم خاص اول یافت شود احتمال وضعیت سوم (نام اثر + نام صاحب اثر) بالا رفته و برچسب HPPL تخصیص می‌یابد که در آن H برای Human، P برای Property (جهت تاکید P دوبار ظاهر شده و این یعنی احتمال برچسب Property بیشتر از بقیه است) و L برای Location بکار رفته است.

در حالت مشابه در صورت وجود کلماتی کلیدی مانند قاره، شهر، کشور، استان، ایالت، روستا، رودخانه، دریا، دریاچه، اقیانوس، فلات، جلگه، کویر، دشت، جزیره، سد و ... احتمال وضعیت دوم (مکان‌های جغرافیایی) بالا می‌رود و برچسب HPLL تخصیص می‌یابد. در صورتی که هیچ کلمهٔ کلیدی در همسایگی این اسامی خاص وجود نداشته باشد که به ما در تعیین نوع این اسامی کمک نماید، برچسب کلّی HPL تخصیص می‌یابد که نشان می‌دهد احتمال رخداد هر یک از سه وضعیت فوق یکسان است.

این سیستم همچنین برای استخراج نمونه‌های موجود در متن و برچسب‌دهی به آنها کلمات هم‌پایه را نیز درنظر می‌گیرد. هنگامی که چند اسم خاص با حروف ربط (‘و’، ‘یا’، ‘،’ و ...) از هم جدا می‌شوند می‌توان آن‌ها را نمونه‌هایی هم‌پایه دانست. در این صورت با برچسب‌دهی به یکی از این اسامی خاص می‌توان برچسب سایر اسامی هم‌پایه را نیز مشخص نمود. در واقع می‌توان گفت این کلمات همگی نمونه‌های یک مفهوم هستند. کلّیهٔ این موارد در متن با برچسب‌های مجزا مشخص می‌شوند.

با انتخاب این برچسب‌های اولّیه برای همهٔ نمونه‌ها فاز برون خط پایان می‌یابد و نمونه‌های برچسب‌دهی شده وارد فاز برخط می‌شوند. در این فاز نیازمند آن هستیم تا برچسب‌های اولیه را به برچسب‌های نهایی که تنها نشانگر یک مفهوم واحد باشند، تبدیل نماییم. در فاز برخط، POPTA از وب و موتور جستجوی گوگل برای رفع ابهام در دسته‌بندی و انتخاب مناسب‌ترین مفهوم از میان مفاهیم کاندیدا و نیز استفاده از نمونه‌های حاشیه‌نویسی شده برای یافتن نمونه‌های دیگر هم‌پایه استفاده می‌نماید. این فرآیند در بخش۴-۳ تشریح شده است. همچنین در صورت مطلوب نبودن نتایج حاصل از بکارگیری گوگل، POPTA با بهره گیری از دائره المعارف ویکی‌پدیا به جستجوی مفهوم متناظر با نمونه مورد بحث می‌پردازد. شرح این روش در بخش ۵-۳ آمده است.

۳-۴- مولفهٔ کار با وب[۲۴]

همانطور که در بخش قبل گفته شد، برای برخی از نمونه‌ها در مرحلهٔ اول لیستی از مفاهیم مرتبط ایجاد می‌شود. برای انتخاب مناسب‌ترین مفهوم از میان این مفاهیم کاندیدا، مجموعه‌ای از پرس و جوها به زبان طبیعی با استفاده از یک سری الگوهای معرفی شده تولید می‌شوند. این الگوها برای زبان فارسی تولید و یا تطبیق یافته‌اند.

نمونه‌ای از پرس و جوهای مناسب برای رفع ابهام برچسب LOC در جدول ۱ آورده شده است که در آن <طبقه> و یا <Category> به یکی از گونه‌های مکان‌های جغرافیایی مانند شهر، استان، کشور، جلگه و ... اشاره دارد.

جدول ۱- چند پرس و جو برای رفع ابهام گونهٔ location برای نمونهٔ X

انگلیسی	فارسی
X and other <Category>s	X و سایر <طبقه>ها
<Category>s such as X	<طبقه>هایی مانند X
...	...
The <Category> of X	<طبقه> X

در این قسمت باید توجّه داشت برخی الگوها مانند الگوهای هیرست، دقت[۲۵] بالایی دارند ولی معمولاً فراخوان[۲۶] آنها بویژه در یک پیکره متنی پایین است. مثلاً الگوهایی مانند "استان‌هایی از جمله ایلام" و یا "جلگه‌هایی مانند آراگون" رخداد بسیار کمی در پیکره‌ها و حتی در سطح وب دارند، از این رو این الگوها برای تعیین نوع نمونه‌هایی که رخداد پایینی در وب دارند، مناسب نیستند. از طرف دیگر الگوهای دیگری (در زبان فارسی) وجود دارند مانند "نام مفهوم + ِ + نام نمونه" (مانند "کشور ایران" و یا "استان ایلام") که فراخوان بسیار بالایی دارند و برای نمونه‌هایی که اسامی آنها کمتر در متون متداول فارسی به چشم می‌خورد (مانند "جلگهٔ آراگون") مناسب هستند ولی در مقابل این الگوها دقت پایینی دارند و عدم توجّه به این مطلب می‌تواند درصد خطا را بالا ببرد. مثلاً یک پرس و جو مانند "تهران بزرگترین شهر ایران است" می‌تواند منجر به درنظر گرفتن "ایران" به عنوان نمونه‌ای از مفهوم "شهر" شود. (این مسئله به دلیل مورد خاص ساختار کسرهٔ اضافه در زبان فارسی رخ می‌دهد و معادلی در زبان انگلیسی ندارد. در فارسی "city of Iran" دقیقاً مانند "Iran city" نوشته می‌شود). بنابراین همواره نیازمند یک مصالحه بین دو معیار دقت و فراخوان خواهیم بود. در واقع در هنگام استفاده از الگوهایی با دقت پایین باید از ویژگی‌های دیگری نیز برای اطمینان از صحت برچسب معنایی انتخابی بهره برد.

بطور کلّی این سیستم ۳ گروه از پرس و جوها را تولید می‌نماید:

۱) پرس و جوهایی که نشان‌دهندهٔ یکی از مکان‌های جغرافیایی می‌باشد؛ مانند [نام نمونه] و سایر [نام مفهوم]ها

۲) پرس و جوهایی که نشان‌دهندهٔ نام یکی از انسان‌ها می‌باشد؛ مانند [لقب] + [نام نمونه]

۳) پرس و جوهایی که نشان‌دهندهٔ نام یکی از محصولات فرهنگی- هنری می‌باشد؛ مانند [نام مفهوم] + -- ِ + [نام نمونه]

در اینجا نیاز به یک مرحله نرمال‌سازی پرس و جوها به منظور انطباق با اطلاعات فارسی موجود در سطح وب داریم. در این مرحله پرس و جوهای مورد نظر باید به ازای همه جایگشت های حروف چندگانه (۲ حالت برای ۲ حرف ی و ک) تولید شوند. این امر می‌تواند حجم پرس و جوها را تا ۴ برابر بالا ببرد و تا حدی از کارایی سیستم می‌کاهد ولی تا هنگامی که کاراکترهای فارسی دارای کدهای استاندارد یکتا نشوند، این سیستم باید بر هرگونه فرمت ورودی منطبق گردد.

پس از تولید پرس و جوهای مناسب، هر یک از طریق Google API در وب جستجو می‌شوند و نتایج بازیابی شده حاصل، توسط برنامه پردازش می‌شود و مفهومی که بالاترین تعداد hit را دارد به عنوان مناسبترین برچسب معنایی انتخاب می‌گردد. این سیستم همچنین هر بار در یک گردش از حلقه، از نمونه‌هایی که نوع آنها تعیین شده برای یافتن نمونه‌های دیگر هم‌پایه در سطح وب و پیکره استفاده می‌کند. برای مثال اگر در مرحلهٔ اول "ایران" به عنوان نمونه‌ای از مفهوم "کشور" انتخاب شود، در مرحلهٔ دوم با استفاده از الگوهایی مانند "ایران، * و سایر کشورها"، کشورهای دیگری (که جایگزین رشتهٔ * می‌شوند) نیز از وب استخراج می‌گردند و یا در سطح پیکره نمونه‌های هم‌پایه دیگر را که با حروف ربط (از جمله 'و'، 'یا'، '،' و ...) به نمونه شناخته شده (در اینجا ایران) مرتبط شده‌اند را شناسایی می‌نماید.

با اتمام این بخش برای بسیاری از نمونه‌های برچسب خورده از مرحلهٔ قبل مناسب‌ترین برچسب معنایی نهایی انتخاب شده و هر نمونه با یک برچسب معنایی به زبان فارسی که اشاره به مفهومی خاص دارد، برچسب‌گذاری می‌گردد. البته هنوز مواردی باقی می‌مانند که در این مرحله نیز برچسب مناسبی برایشان انتخاب نمی‌شود. در این موارد برای برچسب‌گذاری از دائره المعارف ویکی‌پدیا کمک می‌گیریم که شرح فرآیند آن در بخش بعد به تفصیل آمده است.

۳-۵- پردازشگر ویکی‌پدیا

همانطور که گفته شد، پس از اتمام کار فازهای قبل تعدادی از اسامی خاص باقی می‌مانند که بااستفاده از روش‌های مطرح شده قادر به تعیین برچسبی برای آنها نبوده‌ایم.

برای برچسب‌دهی به این اسامی باقیمانده یک راه این است که تمامی برچسب‌های معنایی موجود را به عنوان برچسب‌های کاندیدا در نظر گرفته و سپس با موتور جستجوی گوگل مناسب‌ترین برچسب را که دارای بالاترین تعداد hit می‌باشد برای آن اسم انتخاب نماییم.

از آنجایی که تعداد این برچسب‌های کاندیدا بسیار زیاد است برای انتخاب هر برچسب بطور متوسط ۲۰۰ پرس و جو تولید می‌شود (برای هر نمونه لازم است حدود ۵۰ برچسب کاندیدا و برای هر یک به طور متوسط ۴ پرس و جو آزموده شود تا از میان آنها برچسب مناسب انتخاب گردد). لذا تعیین مناسب‌ترین برچسب برای هر اسم خاص با این روش عملاً غیرممکن است. چرا که این کار کارایی سیستم را بسیار پایین می‌آورد و استفاده از چنین سیستمی را در عمل ناممکن می‌سازد.

یک راه‌حل مناسب برای این مشکل استفاده از ویکی‌پدیا برای تعیین برچسب مناسب می‌باشد. از آنجایی که ویکی‌پدیا یک دایرةالمعارف جامع و فراگیر است که موضوعات متنوع زیادی را پوشش می‌دهد و مطالب آن به بیش از ۱۹۰ زبان مختلف دنیا از جمله زبان فارسی موجود است، می‌تواند منبع مهمی برای تامین اطلاعات مورد نیاز برای بسیاری از کلمات و بویژه اسامی خاص باشد و فعّالیت استخراج دانش را برای سیستم‌های نمونه‌دهی و حاشیه‌نویسی تسهیل نماید و بر کارایی این سیستم‌ها بیفزاید.

در این سیستم در دو مرحله به پردازش اطلاعات بازیابی شده از ویکی‌پدیا می‌پردازیم. در مرحلهٔ اول از ویکی‌پدیا تنها به منظور برچسب‌دهی نمونه‌های موجود در متن استفاده می‌کنیم. در مرحلهٔ دوم که از خروجی‌های مرحلهٔ اول بهره می‌گیرد به استخراج نمونه‌های دیگری که در متن صفحات بازیابی شده وجود دارند، می‌پردازیم.

الف) تعیین مفهوم متناظر با نمونه مستخرج از متن پیکره

در این مرحله کلّیهٔ اسامی خاص موجود در متن که بدون برچسب باقی مانده‌اند به عنوان پرس و جو به ویکی‌پدیا داده می‌شوند و با پردازش جملات اوّلیهٔ صفحات بازیابی شده برچسب مناسب برای آنها تعیین می‌گردد.

با بررسی مقالات موجود در این دانشنامه مشخص می‌گردد که در بیشتر صفحات ویکی‌پدیا که عنوان آن یک اسم خاص است در چند جملهٔ اول متن نام مفهوم متناظر با آن اسم خاص به صراحت ذکر شده است. از آنجایی که در این

سیستم برای تعیین نوع اسامی خاص از ویکی‌پدیا استفاده می‌کنیم، می‌توانیم با یافتن الگوهایی که بیانگر رابطهٔ یک نمونه با مفهوم متناظر آن است، به استخراج مفاهیم دربرگیرندهٔ نمونه‌های ورودی بپردازیم. مثلاً جملاتی مانند "**ایران** کشوری در جنوب غربی آسیا در منطقه‌ای مشهور به خاورمیانه است." و یا "**بینوایان** نام رمان معروفی نوشته ویکتور هوگو نویسنده سرشناس فرانسوی است. " به روشنی بیانگر مفهوم مرتبط با نمونه ورودی است.

اسامی انسان‌ها معمولاً با کلماتی مانند انواع مشاغل (نویسنده، جامعه شناس، کارگردان، بازیگر، فیلمنامه‌نویس و ...)، انواع مسئولیت‌های دولتی - حکومتی (رهبر، رئیس جمهور، نخست وزیر، پیشوا، صدراعظم و ...)، کلماتی که نشاندهندهٔ ارتباط دو یا چند انسان با هم هستند (یاران، صحابه، پدر، مادر، فرزند، همفکری کردن و ...) و سایر کلمات کلیدی که عموماً برای یک انسان بکار می‌روند (شخصیت، رئیس، نوشتهٔ، اثرِ، متولد و ...) هم‌وقوعی دارند. مثلاً:

"**هیتلر** دارای مقاماتی همچون صدر اعظم آلمان، رئیس دولت، و ریاست ایالات، یک دیکتاتور مطلق و سخنرانی با استعداد می‌باشد"، "**نیکی کریمی** متولد ۱۹ آبان ۱۳۵۰ در تهران بازیگر، کارگردان و یک مترجم ایرانی است."، "**وَنسان ویلِم ون گوگ** نقاش نامدار زاده هلند بود."

با پردازش و تجزیه صفحات فارسی ویکی‌پدیا و با توجه به راهنماهایی که در کد این صفحات قرار داده شده‌اند (مانند تعیین اولین رخداد عنوان مقاله در متن – که همان نمونهٔ مورد نظر است –، اولین رخداد هر کلمه‌ای از متن که عنوان صفحهٔ دیگری در ویکی‌پدیاست، کلماتی که هنوز مقالهٔ متناظر با آنها در ویکی‌پدیا ایجاد نشده است و ...) و با پی‌گیری کلمات کلیدی موجود در همسایگی نمونهٔ مورد جستجو (که در مثال‌ها بصورت زیرخط‌دار مشخص شده‌اند) می‌توان به نوع برچسب معنایی متناظر با مفهوم دربرگیرندهٔ آن نمونه پی برد (برای اطلاعات بیشتر در مورد ساختار ویکی‌پدیا و نحوه تجزیه صفحات آن رجوع کنید به [۲]).

در این بخش نیز همانند بخش واسط زبانی نیازمند انجام نرمال‌سازی جهت انطباق برچسب‌های فارسی ورودی و کلمات فارسی موجود در صفحات ویکی‌پدیا می‌باشیم.

اگرچه در بسیاری موارد بررسی کلمات کلیدی موجود در همسایگی نمونه موردنظر مارا به مفهوم متناظر آن می‌رساند، ولی همیشه نمی‌توان نزدیکترین کلمهٔ کلیدی همسایه را به عنوان برچسب معنایی متناظر انتخاب نمود. چرا که (۱) ممکن است بجای کلمه کلیدی نیازمند بررسی عبارات کلیدی باشیم مثلاً اگر فرض کنیم کشور، شهر، استان و ایالت کلمات کلیدی هستند و بدنبال مفهوم متناظر با نمونه "شیراز" یا "تهران" می‌گردیم در جملات "شیراز مرکز استان فارس است" و یا "تهران پایتخت کشور ایران است." وجود کلمهٔ کلیدی استان در همسایگی نمونهٔ "شیراز" و کلمهٔ کلیدی کشور در همسایگی نمونهٔ "تهران"می‌تواند منجر به خطا شود در چنین شرایطی معرفی عبارات کلیدی "پایتخت کشور" و "مرکز استان" که همیشه معادل کلمه کلیدی "شهر" هستند مشکل را حل می‌کند. در POPTA پایگاهی از عبارات کلیدی و کلمات کلیدی معادل آنها معرفی شده که در کاهش خطا بسیار موثر است. واضح است که از آنجا که قادر به تعیین همهٔ عبارات کلیدی موجود در زبان فارسی نیستیم و همچنین به دلیل آن‌که افزایش تعداد کلمات کلیدی موجود میزان خطا در تخصیص نمونه به نزدیکترین کلمهٔ کلیدی را افزایش خواهد داد، لازم است راهکار مکملی برای این مشکل پیشنهاد نماییم.

(۲) ممکن است نزدیکترین کلمه کلیدی همسایهٔ مورد نظر به نمونه دیگری اشاره کند. مثلا در جمله "تهران در کشور ایران واقع شده است" اساساً کلمه کلیدی برای نمونه تهران وجود ندارد و یا در جمله "تهران در کشور ایران بزرگترین شهر است" کلمه کلیدی مناسب برای تهران کلمه شهر است که گرچه در جمله وجود دارد ولی نزدیکترین کلمه کلیدی به نمونه نیست.

در چنین مواردی به عنوان راهکار مکمل در POPTA ۴ کلمهٔ کلیدی نزدیکتر به نمونه را به عنوان مفاهیم کاندیدا درنظر گرفته و سپس با موتور جستجوی گوگل میان این مفاهیم به رفع ابهام پرداخته و مفهومی با بالاترین تعداد hit را به عنوان مفهوم مناسب برای نمونهٔ موردنظر انتخاب می‌نماییم. عدد ۴ به صورت تجربی و پس از آزمون های متعدد بدست آمده است.

ب) یافتن نمونه‌های جدید از متون ویکی‌پدیا

در گام دوم فراتر از نمونه‌دهی هستان‌شناسی مقصد با استفاده از نمونه‌های موجود در متن رفته و به منظور یافتن سایر نمونه‌های موجود در متن و تعیین مفهوم مورد نظر آنها به پردازش کلّی صفحات بازیابی شده و نیز ساختارهای موجود در این صفحات می‌پردازیم.

اسامی خاص موجود در صفحات ویکی‌پدیا که عناوین مقالات دیگری در این دایرةالمعارف هستند، همگی به صفحات مربوط به خود پیوند یافته‌اند. بنابراین می‌توان بسیاری از نمونه‌های موجود در متن و حتی برخی از مفاهیم عمومی‌تر مانند شهر، کشور، زبان، صنعت، انقلاب و ... را استخراج نمود. بیشتر این کلمات پیوند یافته را اسامی خاص تشکیل می‌دهند و مفاهیم عمومی‌تر به ندرت به چشم می‌خورند.

کلمات مهم متن مانند عنوان مقاله (که در متن نیز تکرار می‌شود) و یا عناوین سرفصل‌های یک مقاله نیز می‌توانند در استخراج بخش‌های مهم متن بسیار کمک کننده باشند.

بدلیل ساختار غیر استاندارد ویکی‌پدیا پردازش کلّیهٔ متون صفحه و استخراج اطلاعات مورد نیاز بسیار دشوار است. تاکنون تلاش‌هایی برای توسعهٔ یک واسط برنامهٔ کاربردی (API[۲۷]) برای ویکی‌پدیا صورت گرفته است [۲۶، ۲۷]. بیشتر این APIها برای زبان‌های انگلیسی، ژاپنی و چینی توسعه یافته است. متأسفانه در حال حاضر واسطی برای دسترسی به صفحات فارسی ویکی‌پدیا وجود ندارد.

به دلیل مشکلات موجود برای پردازش و تجزیهٔ صفحات فارسی ویکی‌پدیا توجّه خود را به ساختارهای موجود در این صفحات معطوف کردیم. سیستم POPTA پنج ساختار پیوندها، تغییر مسیر، رده‌ها، ابهام زدایی و جعبه‌های اطلاعاتی را مورد بررسی قرار داده است [۲].

- صفحات تغییر مسیر بیشتر برای تهیهٔ فهرستی از لغات معادل (مترادف، صورت‌های جمع یا مفرد، روش‌های نگارشی مختلف، مخففها و نیز اشتباهات املایی مصطلح) کاربرد دارند و می‌توان در پایان فاز نمونه‌دهی به هستان‌شناسی برای هر نمونه فهرستی از لغات معادل آن را نیز ذخیره نمود.
- پیوندها در ویکی‌پدیا دو نوع هستند: داخلی و خارجی.

(۱) پیوندهای داخلی که به موجودیتی در داخل ویکی‌پدیا – که عموماً عنوان مقاله‌ای در خود این دایرةالمعارف است – اشاره دارند. همانطور که ذکر شد، سیستم POPTA از این پیوندها برای استخراج نمونه‌های موجود در صفحات بازیابی شدهٔ ویکی‌پدیا بهره می‌گیرد. استفادهٔ دیگر این پیوندها ایجاد مجموعهٔ لغات هم‌وقوع با عنوان صفحه و نیز یافتن موجودیت‌های مرتبط با نمونهٔ موردنظر است؛ ساختار مشابه دیگر "جستارهای وابسته" است که به مفاهیم و موجودیت‌های مرتبط با نمونهٔ جستجو شده اشاره دارد.

(۲) پیوندهای خارجی (یا "پیوند به بیرون") که به صفحاتی در خارج از سایت ویکی‌پدیا اشاره دارند. این سیستم در حال حاضر به این دسته از پیوندها نپرداخته است.

- رده‌های موجود در ویکی‌پدیا ساختارهای بسیار مفیدی برای تعیین حوزهٔ دربرگیرندهٔ مقالهٔ بازیابی شده است و هنگامی که این مقاله یک اسم خاص را توصیف می‌کند، این رده‌ها می‌توانند تعیین کنندهٔ مفهوم متناظر با آن نمونه باشند. خوشبختانه این رده‌ها و ساختار پیوندهای آنها بصورت جدول‌های SQL موجودند. از این رو می‌توان بدون تجزیه و پردازش مقالات موجود در ویکی‌پدیا به

آنها دسترسی داشت. پیاده‌سازی این بخش نیز در فهرست کارهای آیندهٔ این سیستم قرار می‌گیرد.

- ساختارهای ابهام‌زدایی نیز ساختارهای بسیار مفیدی برای رفع ابهام معنایی لغات و یافتن معانی مختلف هر کلمه[۲۸] (WSD) است. این ویژگی ویکی‌پدیا برای بخش WSD حاشیه‌نویسی می‌تواند بکار رود. در فرایند WSD یا رفع ابهام سعی می شود تشخیص داده شود از میان معانی متعدد یک کلمه کدام معنی در رخداد حاضر آن مدنظر بوده است. یافتن مجموعه لغات هم‌وقوع با هر یک از معانی نمونهٔ مورد نظر (با استفاده از پیوندهای موجود در متن مقالهٔ مربوط به هر یک از این نمودها) و محاسبهٔ میزان اشتراک و هم‌پوشانی این لغات با لغات موجود در همسایگی نمونهٔ استخراج شده از متن می‌تواند به تعیین معنی نمونهٔ استخراج شده کمک نماید. در مرحلهٔ بعد می‌توان مفهوم متناظر با هر معنی نمونهٔ جستجو شده را تعیین نمود.
- جعبه‌های اطلاعاتی یا Infoboxها از پرکاربردترین ساختارهای موجود در ویکی‌پدیا می‌باشند. Infoboxها کلّیهٔ اطلاعات مهم موجود در متن را بصورت یک جدول خلاصه می‌سازند. در میان Infoboxهای موجود برای انواع مقالات در ویکی‌پدیا، اطلاعات موجود در جداول صفحات مربوط به کشورها، شهرها و انسان‌ها اهمیّت بیشتری دارند. سایر موجودیت‌های موجود در ویکی‌پدیا معمولاً فاقد جدول می‌باشند.

درحال حاضر سیستم POPTA اطلاعات موجود در Infoboxهای مربوط به کشورها و شهرها را استخراج می‌نماید. پیاده‌سازی استخراج اطلاعات جداول موجود در صفحات مربوط به شخصیت‌های انسانی در فهرست کارهای آیندهٔ این سیستم قرار می‌گیرد.

۳-۶- واسط میان زبانی

همانطور که پیش از این گفته شد، POPTA یک سیستم میان زبانی است. دلیل این نامگذاری آن است که این سیستم متون فارسی را با استفاده از یک هستان‌شناسی انگلیسی حاشیه‌نویسی کرده و هستان‌شناسی انگلیسی را نیز با استفاده از متون فارسی نمونه‌دهی می‌نماید. بنابراین نیازمند واسطی هستیم تا بتواند برچسب‌های انتخاب شده را که به زبان فارسی هستند، به مفهوم معادل آنها در هستان‌شناسی ورودی که به زبان انگلیسی است، متصل نماید.

می‌دانیم که هر کلمهٔ فارسی می‌تواند معادل چندین کلمهٔ انگلیسی باشد و برعکس. در نتیجه بسیار محتمل است که هیچ تناظر یک به یکی بین مفاهیم انگلیسی موجود در هستان‌شناسی و مفاهیم فارسی مرتبط با نمونه‌های انتخاب شده وجود نداشته باشد.

برای یافتن مناسب‌ترین تناظر، نیازمند اعمال روش‌های رفع ابهام معنایی کلمات (WSD) هستیم. به این منظور دو راه حل وجود دارد: ۱) ترجمهٔ دستی بخشی از مجموعه‌های ترادف[۲۹] وردنت و اتصال خودکار مفاهیم انتخاب شدهٔ فارسی به این مجموعه‌های ترادف ترجمه شده با استفاده از ترکیبی از معیارهای شباهت زبانی (مانند معیار edit-distance) و ساختاری، ۲) ترجمهٔ خودکار مفاهیم انتخاب شده توسط یک دیکشنری دو زبانه و انتخاب مجموعه‌های ترادف معادل در وردنت.

مشکل ابهامات معنایی به انتخاب معنی[۳۰] مناسب کلمه (یا عبارت) نمونه و نیز انتخاب ترجمهٔ مناسب آن (مفهوم متناظر) در زبان مقصد (انگلیسی) توجه دارد و در بخش حاشیه‌نویسی جدی‌تر از بخش نمونه‌دهی است. برای رفع ابهام در ترجمه و انتخاب معنی، یک رویکرد جدید مکاشفه‌ای مبتنی بر دیکشنری معرفی نموده‌ایم.

در این رویکرد به منظور ترجمهٔ برچسب‌های فارسی به زبان انگلیسی در مرحلهٔ اول از یک دیکشنری فارسی به انگلیسی که بصورت یک پایگاه دادهٔ دو زبانه موجود است، بهره بردیم. پس از انجام نرمال سازی برچسب‌ها جهت انطباق با فرمت دیکشنری، با دادن هر کلمهٔ فارسی به عنوان ورودی به این دیکشنری مجموعه‌ای از لغات انگلیسی معادل بازیابی می‌شوند. این مجموعه از لغات بصورت الفبایی مرتب هستند؛ بنابراین ممکن است برخی از این لغات ترجمه‌های مصطلح واژهٔ فارسی ورودی نبوده و وجود آنها در لیست لغات انگلیسی می‌تواند تعداد مفاهیم معادل انتخاب شده در هستان‌شناسی مقصد شده را افزایش دهد و این امر منجر به افزایش ابهام در نمونه‌دهی خواهد شد.

برای حل این مشکل این بار در جهت عکس عمل کرده و کلّیهٔ لغات انگلیسی بازیابی شده از دیکشنری اول را به یک دیکشنری انگلیسی به فارسی دیگر می‌دهیم تا برای هر کلمهٔ انگلیسی لغات معادل آن به زبان فارسی بدست آیند. برای این کار نمی‌توان از نسخهٔ انگلیسی به فارسی همان دیکشنری اول استفاده نمود؛ چرا که همان مجموعه لغات فارسی اوّلیه بدست می‌آیند و این امر کمکی به ما نمی‌کند. بنابراین در این بخش از دیکشنری انگلیسی به فارسی آنلاین آریان‌پور استفاده کردیم (که در آن معانی به ترتیب مصطلح بودن مرتب شده‌اند) و برای هر لغت انگلیسی بدست آمده از دیکشنری قبلی مجموعهٔ واژه‌های معادل آن را در زبان فارسی بازیابی نمودیم.

نکتهٔ قابل توجّه این است که همانطور که گفته شد این بار این مجموعه لغات بازیابی شده به ترتیب مصطلح بودن و ارتباط بیشتر با مفهوم مورد نظر مرتب هستند و این امر به ما در انتخاب بهترین ترجمه‌های معادل از میان این مجموعه لغات بازیابی شده کمک می‌نماید.

در این مرحله برای هر لغت انگلیسی بدست آمده از ترجمهٔ اول مجموعه‌ای از لغات فارسی معادل داریم که از ترجمهٔ دوم بدست آمده و به ترتیب مربوط بودن و کاربرد مرتب هستند. حال کافیست از بین این کلمات انگلیسی اولین کلمه‌ای را انتخاب نماییم که کلمهٔ فارسی اوّلیه (برچسب معنایی ورودی) در میان کلمات فارسی معادل آن کلمهٔ انگلیسی زودتر رخ داده باشد (در اولویت بالاتری قرار گرفته باشد). در واقع معمولاً کلمه‌ای مناسب است که کلمهٔ فارسی اوّلیه، اوّلین معادل آن باشد و در صورتی که این کلمهٔ فارسی اولین معادل هیچ یک از کلمات انگلیسی بدست آمده از ترجمهٔ اول نباشد به سراغ معادل‌های بعدی می‌رویم و در صورتی که این کلمه برای بیش از یک کلمهٔ انگلیسی در اولین مکان واقع شده باشد، همهٔ آن لغات انگلیسی انتخاب خواهند شد.

برای روشن‌تر شدن موضوع برچسب کشور را در نظر بگیرید. در مرحلهٔ اول این کلمه به دیکشنری فارسی به انگلیسی ذکر شده داده می‌شود. کلمات معادل انگلیسی زیر به ترتیب الفبایی بدست می‌آیند:

Commonwealth, Country, Kingdom, Nation, Soil, State, Territory

حال هریک از این کلمات را به دیکشنری دوم می‌دهیم، نتایج حاصل در جدول ۲ نمایش داده شده‌اند.

جدول ۲- نتایج حاصل از ترجمهٔ لغت "کشور" با استفاده از دو دیکشنری

فارسی	انگلیسی
-	Commonwealth
کشور، دیار، بیرون شهر، دهات، ییلاق	Country
پادشاهی، کشور، قلمرو پادشاهی	Kingdom
ملّت، قوم، امّت، خانواده، طایفه، کشور	Nation
خاک، کثیف کردن، ....، زمین، کشور، سرزمین،	Soil
حالت، کشور، ایالت	State
سرزمین، خاک، خطّه، زمین، ملک، کشور، قلمرو	Territory

همانطور که دیده می شود، کلمه کشور به عنوان اولین ترجمه کلمه Country، دومین ترجمه کلمات Kingdom و State، ششمین ترجمه کلمات Nation و Territory و ... رخ داده است. بنابراین کلمهٔ "Country" به عنوان مناسب‌ترین ترجمه برای برچسب فارسی *کشور* انتخاب می‌گردد. نتایج ارزیابی این روش در بخش ۵ آورده شده است.

۳-۷- نمونه‌دهی به هستان‌شناسی

پس از ترجمهٔ کلّیهٔ برچسب‌های معنایی بدست آمده از مرحلهٔ برچسب‌زنی معنایی نوبت به یافتن مفهوم معادل هر برچسب (ترجمه شده) در هستان‌شناسی مقصد است.

در حال حاضر هستان‌شناسی وردنت به عنوان هستان‌شناسی مقصد سیستم POPTA انتخاب شده است و مفاهیم آن با استفاده از نمونه‌های موجود در متن نمونه‌دهی می‌شوند. با توجه به وجود نگاشت هایی میان وردنت و هستان‌شناسی های دیگر (مثلا SUMO) امکان برقراری اتصال میان متون فارسی و سایر هستان‌شناسی ها نیز وجود خواهد داشت.

از آنجایی که هر مفهوم در هستان‌شناسی وردنت از مجموعه‌ای از لغات هم‌معنی تشکیل شده است، عملیات یافتن مفهوم معادل برچسب مورد نظر تسهیل می‌گردد و نیازی به درنظر گرفتن معادل‌های این برچسب برای تطابق با نام یکی از مفاهیم هستان‌شناسی وجود ندارد. از سوی دیگر یک کلمه بسته به معانی مختلفی که می‌تواند داشته باشد ممکن است در بیش از یک مجموعهٔ ترادف ظاهر شود. این امر موجب می‌شود برای هر نمونه بیش از یک مفهوم متناظر در وردنت وجود داشته باشد و لذا مجددا نیاز به روش های WSD خواهیم داشت. برای مثال با استفاده از واسط میان زبانی، معادل انگلیسی issue برای مفهوم "نشریه" انتخاب می‌گردد. با توجّه به معانی مختلف این کلمه دو مجموعهٔ ترادف متناظر زیر در وردنت انتخاب می‌شوند:

- an important question that is in dispute and must be settled;
- one of a series published periodically;

حال آنکه تنها مجموعه ترادف دوم مورد نظر ماست.

در حالت عمومی می‌توان برای هر یک از برچسب‌های معنایی موجود در سیستم یک تعریف زبان طبیعی درنظر گرفت. برای یافتن این تعاریف زبان طبیعی می‌توان از دایرةالمعارف آزاد ویکی‌پدیا بهره برد. اغلب هنگامی که یک مفهوم عمومی در ویکی‌پدیا جستجو می‌شود، صفحهٔ بازیابی شده حاوی یک تعریف زبان طبیعی کوتاه و کامل برای آن مفهوم خواهد بود. به علاوه می‌توان لیست کلمات هم‌وقوع با این مفهوم را نیز با تجزیهٔ صفحهٔ بازیابی شده و پردازش کلماتی که به صفحات دیگر این دایرةالمعارف پیوند یافته‌اند، استخراج نمود.

با داشتن یک تعریف زبان طبیعی برای هر برچسب و یا کلمات هم‌وقوع با آن برچسب و ترجمهٔ آنها به زبان انگلیسی می‌توان با توجّه به شباهت معنایی میان تعریف هر برچسب با تعریف بازیابی شده از وردنت برای هر مجموعهٔ ترادف و میزان هم‌پوشانی این دو تعریف، مناسب‌ترین مجموعهٔ ترادف را از میان مجموعه‌های ترادف کاندیدا انتخاب نمود. مثلاً برای مفهوم کتاب تعاریف زیر ارائه شده:

- ویکی‌پدیا:

"**کتاب** مجموعه‌ای از صفحه‌های کاغذی است که متنی روی آن‌ها نوشته شده‌است و در یکی از طرف‌ها به هم بسته شده‌اند و مجلد است."

- وردنت:

"a written work or composition that has been published"
"physical objects consisting of a number of pages bound together"

در اینجا باید توجّه داشت برچسب‌هایی که در حوزهٔ کار این سیستم قرار دارند عموماً دارای معانی متفاوتی نیستند و می‌توان گفت مصطلح‌ترین معنی هر یک از این برچسب‌ها همان معنی مورد نظر است.

از آنجایی که وردنت برای هرکلمه عددی نگه می‌دارد که نشان‌دهندهٔ فرکانس ارجاع آن کلمه به یک مجموعهٔ ترادف خاص است [۲۸]، درواقع می‌توان مجموعهٔ ترادفی که بالاترین فرکانس ارجاع را دارد مصطلح‌ترین و پرکاربردترین مجموعهٔ ترادف برای کلمهٔ ورودی دانست و مجموعهٔ ترادف با بالاترین فرکانس ارجاع را به عنوان مجموعهٔ ترادف مناسب برای برچسب معنایی مورد نظر انتخاب نمود. در صورت وجود بیش از یک مجموعهٔ ترادف با ماکسیمم فرکانس ارجاع همهٔ این مجموعه‌های ترادف به عنوان مفهوم معادل نمونهٔ موردنظر انتخاب می‌گردند. برای مثال برای برچسب سد معادل انگلیسی dam انتخاب می‌گردد که در وردنت دارای سه مجموعه ترادف متناظر زیر می‌باشد:

- a barrier constructed to contain the flow of water or to keep out the sea
- a metric unit of length equal to ten meters
- female parent of an animal especially domestic livestock

از آنجایی که فرکانس ارجاع به مجموعه ترادف اول بیشتر از دو مجموعهٔ دیگر است، این مجموعه به عنوان مفهوم متناظر در وردنت انتخاب شده و نمونه‌های استخراج شده به آن متصل می‌گردند.

نتایج ارزیابی این روش انتخاب مجموعه ترادف در بخش ۵ آورده شده است.

به این ترتیب در پایان مرحلهٔ نمونه‌دهی، هر نمونه به یک یا چند مفهوم معادل که می‌توانند معادل یک یا چند مجموعهٔ ترادف از وردنت باشند، متصل می‌گردد.

۳-۸- حاشیه‌نویسی معنایی

همانطور که مطرح شد، در این سیستم حاشیه‌نویسی معنایی همزمان با نمونه‌دهی به هستان‌شناسی انجام می‌گیرد. به عبارت دیگر هنگامی که مفهوم معادل نمونهٔ استخراج شده از متن مشخص گردید و محل آن در هستان‌شناسی ورودی تعیین شد، بخشی از متن که دربرگیرندهٔ آن نمونه است با برچسبی همنام با شمارهٔ مفهوم معادل در هستان‌شناسی حاشیه‌نویسی می‌شود.

در این مرحله برای هر نمونهٔ موجود در متن، برچسب معنایی تعیین شده، شمارهٔ مجموعهٔ ترادف (یا مجموعه‌های ترادف) معادل در وردنت و تعریف زبان طبیعی این مجموعه‌های ترادف ذخیره می‌شود.

بنابراین تا این جا از یک پیکرهٔ فارسی دارای برچسب‌های مقولهٔ نحوی برای نمونه‌دهی به هستان‌شناسی و نیز حاشیه‌نویسی معنایی همان پیکره بهره بردیم. در مرحلهٔ بعد می‌توانیم از همین پیکره که در حال حاضر برچسب‌های معنایی را نیز علاوه بر برچسب‌های مقولهٔ نحوی داراست، برای حاشیه‌نویسی معنایی اسناد موجود در وب استفاده نماییم.

یک راه انجام این کار بکارگیری عملیات یادگیری ماشینی است تا با ترکیب روش‌های آماری و مبتنی بر الگو، قوانین موجود در این پیکره را بیاموزد و با کمک این قوانین جدید و اضافه نمودن آنها به مجموعه قوانین مکاشفه‌ای قبلی به حاشیه‌نویسی معنایی اسناد موجود بر روی وب بپردازد.

۴- آزمون و ارزیابی

این سیستم برای ارزیابی نتایج حاصل از فاز استخراج برچسب‌های معنایی و فاز نمونه‌دهی به هستان‌شناسی، بر روی ۱۵ متن از متون پیکرهٔ فارسی موجود - شامل ۸۰۸۷۹ کلمه - آزمایش گردید و دو معیار دقت و فراخوان برای دو

پیمانه نمونه‌دهی به هستان‌شناسی و حاشیه‌نویسی معنایی در هر فاز بطور جداگانه محاسبه گردید. در این بخش نتایج ارزیابی را برای سه بخش (الف) انتخاب برچسب معنایی یا مفهوم متناظر با نمونه‌ها، (ب) ترجمه صحیح مفهوم موردنظر به انگلیسی و (ج) یافتن مفهوم (مجموعه ترادف) متناظر با مفهوم یافت شده در وردنت ارائه نموده‌ایم.

الف) انتخاب برچسب معنایی یا مفهوم متناظر با نمونه‌ها

جدول ۳ نتایج حاصل از ارزیابی سیستم (۱) تولید برچسب‌های معنایی – بدون استفاده از ویکی‌پدیا و (۲) تولید برچسب‌های معنایی – پس از استفاده از ویکی‌پدیا را نشان می‌دهد.

همانطور که در جدول ۳ فاز اول مشاهده می‌شود این سیستم از دقت بالایی برخوردار است. بیشتر خطاهای رخ داده را می‌توان به ۲ دستهٔ کلّی تقسیم‌بندی نمود:

- خطاهای ناشی از استفاده از الگوهایی به شکل "نام مفهوم + اضافه + نام نمونه" که همانطور که قبلاً گفته شد فراخوان بالا ولی دقت پایین دارند. مثلاً در عبارتی مانند "یزد، شهر ادیبان و عارفان" با استفاده از این الگو کلمهٔ "ادیبان" را به عنوان نمونه‌ای از مفهوم "شهر" استخراج می‌کند.

این دسته از خطاها را می‌توان با استفاده از ویژگی‌های دیگری که برای هر مفهوم قابل استخراج است، کاهش داد. این ویژگی‌ها می‌توانند از توصیف‌های زبان طبیعی موجود برای هر مفهوم در هستان‌شناسی ورودی و یا در سطح وب و نبز از منابع اطلاعاتی دیگر مانند انواع دایرةالمعارف‌های موجود در سطح وب (از جمله ویکی‌پدیا) استخراج شوند. راه‌حل دیگر آن است که از دایرةالمعارف ویکی‌پدیا علاوه بر برچسب‌دهی نمونه‌های باقیمانده برای بررسی میزان صحت نتایج حاصل از جستجوی گوگل بهره ببریم. بدین منظور می‌توان نتایجی که از جستجوی پرس و جوهایی با الگوی ذکر شده بدست آمده را بار دیگر در ویکی‌پدیا جستجو نماییم تا درستی یا نادرستی آن مشخص گردد.

- خطاهای ناشی از WSD که تنها بربخش حاشیه‌نویسی تاثیر می‌گذارند. این خطاها به دلیل عملیات رفع ابهام برخی کلمات توسط موتور جستجوی گوگل رخ می‌دهند. از آنجایی که گاه این جستجو مستقل از متن ورودی و در سطح وب صورت می‌گیرد و نمونهٔ موجود در متن به مفهومی با بیشترین تعداد رخداد در سطح وب تخصیص می‌یابد، علیرغم این‌که مفهوم انتخاب شده مفهومی مناسب برای نمونهٔ استخراج شده است و خطایی در بخش نمونه‌دهی به هستان‌شناسی رخ نداده است، ممکن است اتصال این مفهوم به آن بخش از متن صحیح نباشد؛ چرا که نمونهٔ موجود در متن می‌تواند معنی دیگری از مفهوم یافت شده در وب باشد. در این بخش نیز می‌توان از دایرةالمعارف ویکی‌پدیا جهت رفع ابهام معنایی این نمونه‌ها بهره برد. با توجه به ویژگی‌ها و امکانات منحصر بفرد این دانشنامه و بهره‌برداری سیستم‌های مختلف از آن به‌منظور استخراج دانش و معنا می‌توان استفاده از این دایرةالمعارف را جهت ارتقاء بخش حاشیه‌نویسی در فهرست کارهای آیندهٔ این سیستم قرار داد.

همانطور که در جدول ۳ فاز دوم مشاهده می‌شود، استفاده از ویکی‌پدیا به میزان قابل توجهی فراخوان سیستم را افزایش داده و بر کارایی سیستم افزوده است. میزان کاهش دقت سیستم در این فاز در مقابل این افزایش قابل چشم‌پوشی است.

تعداد خطاهای این فاز کمتر از فاز قبل است. خطاهای موجود در این فاز عمدتاً به ناکامل بودن مجموعهٔ کلمات کلیدی و مکاشفه‌های بکار رفته مربوط می‌شود. با گسترش مجموعهٔ لغات هم‌وقوع با هر یک از برچسب‌ها و افزایش مکاشفه‌های مناسب برای یافتن مرتبط‌ترین کلمهٔ کلیدی همسایه می‌تواند میزان خطاهای این بخش را کاهش دهد.

از سوی دیگر از آنجایی که سایز متن‌های مختلف مورد استفاده در آزمون‌ها و درنتیجه میزان رخداد کلمات کلیدی و نسبت آن ها به کل متن متغیر بوده است، برآن شدیم تا نتایج ارزیابی را یکبار به صورت میانگین عادی و یک بار به صورت میانگین وزندار محاسبه کنیم. در میانگین‌گیری وزندار به هر سند وزنی متناسب با حجم آن تخصیص داده‌ایم. در ادامه در ارائه نتایج ارزیابی‌ها میانگین‌های وزندار داخل پرانتز و در مقابل میانگین‌های عادی نوشته شده‌اند. قابل مشاهده است که میانگین‌گیری وزندار تغییر عمده‌ای در میزان دقت و فراخوان ایجاد نمی‌کند

جدول ۳- نتایج ارزیابی تولید برچسب‌های معنایی در سیستم POPTA

۱) فاز اول- بدون استفاده از ویکی‌پدیا

ورودی	نمونه‌دهی به هستان‌شناسی		حاشیه‌نویسی معنایی	
	دقت	فراخوان	دقت	فراخوان
۱	٪۹۵	٪۷۶	٪۹۹	٪۹۲
۲	٪۹۷	٪۶۳	٪۹۹	٪۹۱
۳	٪۹۱	٪۵۱	٪۸۹	٪۳۶
۴	٪۱۰۰	٪۶۰	٪۱۰۰	٪۵۲
۵	٪۸۲	٪۵۶	٪۸۰	٪۷۷
۶	٪۱۰۰	٪۶۴	٪۱۰۰	٪۸۷
۷	٪۱۰۰	٪۶۹	٪۱۰۰	٪۶۹
۸	٪۸۹	٪۸۶	٪۹۲	٪۹۳
۹	٪۹۴	٪۵۲	٪۹۴	٪۹۰
۱۰	٪۹۵	٪۴۳	٪۹۶	٪۳۱
۱۱	٪۱۰۰	٪۵۲	٪۱۰۰	٪۸۰
۱۲	٪۸۱	٪۶۳	٪۵۶	٪۳۰
۱۳	٪۹۲	٪۳۸	٪۹۳	٪۳۴
۱۴	٪۸۷	٪۴۶	٪۹۰	٪۱۶
۱۵	٪۹۳	٪۵۵	٪۹۷	٪۷۴

۲) فاز دوم- با استفاده از ویکی‌پدیا

ورودی	نمونه‌دهی به هستان‌شناسی		حاشیه‌نویسی معنایی	
	دقت	فراخوان	دقت	فراخوان
۱	٪۹۶	٪۹۷	٪۹۵	٪۹۷
۲	٪۹۳	٪۹۲	٪۹۴	٪۹۲
۳	٪۸۸	٪۹۲	٪۸۵	٪۵۰
۴	٪۹۷	٪۹۴	٪۹۸	٪۸۷
۵	٪۸۶	٪۷۵	٪۸۵	٪۹۶
۶	٪۹۵	٪۹۳	٪۱۰۰	٪۸۷
۷	٪۹۷	٪۹۱	٪۹۵	٪۹۵
۸	٪۹۰	٪۹۳	٪۹۴	٪۹۶
۹	٪۹۷	٪۸۷	٪۹۶	٪۹۷
۱۰	٪۹۳	٪۸۳	٪۹۳	٪۸۲
۱۱	٪۱۰۰	٪۹۴	٪۹۸	٪۸۷
۱۲	٪۸۷	٪۹۴	٪۸۹	٪۹۲
۱۳	٪۹۴	٪۸۱	٪۹۷	٪۹۵
۱۴	٪۹۲	٪۶۳	٪۹۶	٪۵۸
۱۵	٪۹۳	٪۷۰	٪۹۷	٪۸۸

بطور میانگین این سیستم قبل از استفاده از ویکی‌پدیا به دقت ۹۳٪ (۹۳/۴) و فراخوان ۵۸/۲٪ (۵۶/۱) در بخش نمونه‌دهی به هستان‌شناسی و دقت ۹۲/۳٪ (۹۲/۴) و فراخوان ۶۳/۴٪ (۵۹) در بخش حاشیه‌نویسی معنایی دست یافته است. خروجی نهایی سیستم که پس از استفاده از ویکی‌پدیا بدست آمده است دقت ۹۳/۲٪ (۹۳/۲) و فراخوان ۸۶/۶٪ (۸۷) در بخش نمونه‌دهی به هستان‌شناسی و دقت۹۴٪ (۹۴) و فراخوان ۸۶/۶٪ (۸۶/۲) در بخش حاشیه‌نویسی معنایی را نشان می‌دهد.

مطابق نتایج فوق استفاده از ویکی پدیا منجر به افزایش کارایی سیستم با حذف موارد نامعلوم در فازهای قبلی می شود. برای مثال اسامی خاصی که هیچ کلمهٔ کلیدی مناسبی در همسایگی آنها قرار ندارند و مکاشفه‌ای نیز برای تشخیص نوع آنها در دست نداریم در فاز اول بدون برچسب باقی می‌مانند. مثلاً در جملهٔ "در آن هنگام بوشهر مرکز عبور بازرگانانی بود که از هند و هندوچین و چین به سوی عمان و آفریقا و اروپا کالا حمل می کردند"، کلمات "بوشهر"، "هند"، "هندوچین"، "چین"، "عمان"، "آفریقا" و "اروپا" که همگی اسامی خاص هستند در فاز اول بدون برچسب باقی می‌مانند؛ حال آن‌که با دادن هر یک از این کلمات به عنوان پرس و جو به ویکی‌پدیا به ترتیب برچسب‌های "شهر"، "کشور"، "منطقه"، "کشور"، "کشور"، "قاره" و "قاره" به این هفت نمونه تخصیص می‌یابد.

البته در این مرحله نیز به علت وجود کلماتی با معانی گوناگون، ممکن است در بخش حاشیه‌نویسی دچار مشکل شویم، در حالی که عملیات نمونه‌دهی به خوبی صورت می‌گیرد. .مثلاً در جملهٔ "داستان‌های کهن ایران را از پهلوی به فارسی دری ترجمه کرده‌اند. " کلمات "ایران"، "پهلوی" و "فارسی" در فاز اول بدون برچسب باقی می مانند. پس از جستجو در ویکی‌پدیا به ترتیب برچسب‌های "کشور"، "دودمان" و "زبان" به این سه نمونه تخصیص می‌یابد. حال آنکه مفهوم متناظر با نمونهٔ پهلوی، در این جمله، "زبان" است و در این مرحله با خطای بخش حاشیه‌نویسی روبرو هستیم (گرچه در بخش نمونه‌دهی خطایی نداریم).

کلماتی که عنوان هیچ مقاله‌ای در ویکی‌پدیا نیستند (مانند "بهارستان") و یا در صفحهٔ مربوط به آنها کلمهٔ کلیدی مناسبی به کار نرفته است (مانند "خورشید")، در انتهای فاز دوم نیز بدون برچسب باقی می‌مانند.

ب) ترجمه صحیح مفهوم موردنظر به انگلیسی

با ارزیابی روش مطرح شده در بخش ۶-۳ مشاهده گردید که این روش برای ۹۶.۲٪ برچسب‌های فارسی معادل انگلیسی مناسبی ارائه داده است.

ج) یافتن مفهوم (مجموعه ترادف) متناظر با مفهوم یافت شده در وردنت

پس از انتخاب ترجمهٔ مناسب برای هر برچسب به مرحلهٔ یافتن مفهوم متناظر با هر برچسب در هستان‌شناسی وردنت می‌رسیم. با ارزیابی نتایج حاصل از روش مطرح شده در بخش ۷-۳ مشاهده گردید برای ۸۱٪ برچسب‌ها، مناسب‌ترین مفهوم همان پرارجاع‌ترین مفهوم می‌باشد؛ برای ۹/۲٪ از برچسب‌ها پرارجاع‌ترین مفهوم گرچه مناسب است ولی مناسب‌ترین مفهوم نیست و لذا خطایی نیز رخ نداده است و برای ۹/۸٪ از برچسب‌ها انتخاب پرارجاع‌ترین مفهوم ایجاد خطا می‌نماید.

۵- نتیجه‌گیری

در این مقاله سیستم POPTA را به عنوان نخستین سیستم نمونه‌دهی و حاشیه‌نویسی موازی برای متون فارسی معرفی نمودیم. POPTA با هدف ایجاد و توسعهٔ سیستم‌های مورد نیاز با استفاده از منابع اولّیهٔ محدود ایجاد شده است. این سیستم هستان‌شناسی‌های انگلیسی را با استفاده از متون فارسی نمونه‌دهی نموده و متون فارسی را نیز با توجه به این هستان‌شناسی حاشیه‌نویسی می‌نماید. یکی از خروجی‌های POPTA یک پیکرهٔ فارسی حاشیه‌نویسی شده برای انجام حاشیه‌نویسی‌های معنایی بیشتر مبتنی بر پیکره برای زبان فارسی خواهد بود.

POPTA در حال حاضر برای نمونه‌دهی هستان‌شناسی وردنت با در نظر گرفتن اسامی خاص به عنوان نمونه‌هایی از مفاهیم این هستان‌شناسی پیاده‌سازی شده است. گام بعدی برای توسعهٔ این سیستم در نظر گرفتن اسامی عام و افعال به عنوان نمونه‌ها و نیز اتصال آنها به هستان‌شناسی‌های دیگر از جمله SUMO می‌باشد.

ارتقاء مولّفهٔ WSD نیز یکی دیگر از بخش‌هایی است که در لیست فعالیت‌های آینده قرار دارد.

برای پیشرفت هرچه بیشتر سیستم‌های استخراج‌گر دانش و حرکت به سوی داشتن یک وب معنایی برای زبان فارسی، در مرحلهٔ اول نیازمند توسعهٔ مولفه‌های آمادهٔ پردازش زبان فارسی و نیز مولفه‌های استخراج اطلاعات (IE) می‌باشیم. این امر نمونه‌دهی به هستان‌شناسی‌ها و نیز حاشیه‌نویسی معنایی متون فارسی را تسهیل می‌نماید.

ایجاد استاندارد یکتا برای کدگزاری‌های مربوط به کاراکترهای زبان فارسی نیز می‌تواند به اعتبار اطلاعات استخراج شده از سطح وب بیفزاید.

استفاده از وب و Google API کمک شایانی به عملیات رفع ابهام در دسته‌بندی و غلبه بر گلوگاه استخراج دانش کرده است ولی مشکلاتی نیز در این میان وجود دارد. از قبیل اطلاعات نادرست موجود در سطح وب که می‌تواند منجر به استخراج دانش نادرست از متون گردد و نیز غیر قابل اعتماد بودن گوگل که گاهی باعث می‌گردد نتایج حاصل از اجرای چند بارهٔ سیستم با ورودی‌های یکسان، متفاوت باشد. برای رفع این مشکل نیازمند اجرای چند بارهٔ سیستم برای هر متن ورودی می‌باشیم.

همانطور که در بخش ارزیابی دیدیم، استفاده از ویکی‌پدیا به میزان قابل توجهی فراخوان سیستم را افزایش داده و بر کارایی سیستم افزوده است. به علاوه خطاهای این فاز نیز از فاز قبل بسیار کمتر و سرعت آن نیز بالاتر است. بنابراین مهمترین کاری که در فهرست کارهای آیندهٔ این سیستم قرار می‌گیرد افزایش نقش ویکی‌پدیا در برچسب‌دهی به نمونه‌های استخراج شده از متن و همچنین بررسی میزان درستی برچسب‌های تخصیص داده شده در فاز اول می‌باشد.

برای گسترش بخش ویکی‌پدیا در این پروژه نیازمند بهبود بخشی به روش‌های استخراج اطلاعات از صفحات این دایرةالمعارف و نیز آشنایی بیشتر با ساختارهای موجود در آن هستیم. بنابراین در اولین گام به توسعهٔ یک تجزیه‌گر صفحات ویکی‌پدیا پرداخته و در گام دوم از ساختارهای موجود در این دایرةالمعارف برای استخراج اطلاعات معنایی در حوزه‌ای گسترده‌تر بهره می‌گیریم.

مراجع

[1] L. Reeve, and H. Han, "Survey of Semantic Annotation Platforms," *Proc, ACM Symp. on Applied Computing,* PP. 1634-1638, 2005.

[۲] ب. صراف‌زاده، *نمونه‌دهی هستان‌شناسی و حاشیه‌نویسی خودکار متون فارسی*، پایان‌نامه کارشناسی، دانشکده مهندسی برق و کامپیوتر، دانشگاه شهید بهشتی، ۱۳۸۷.

[3] M. Vargas-Vera, and D. Celjuska, "Event Recognition on News Stories and Semi-Automatic Population of an Ontology," *Proc, IEEE/WIC/ACM Inte'l Conf. on Web Intelligence*, pp. 615-618, 2004.

Learning And Population: Bridging The Gap Between Text And Knowledge, pp. 26-32, 2006.

[17] V. Borkar, K. Deshmukhy, and S. Sarawagiz, "Automatic Segmentation of Text into Structured Records," *Proc, ACM SIGMOD Int'l Conf. on Management of Data*, pp. 175-186, 2001.

[18] M. Vargas-Vera, E. Motta, J. Domingue, M. Lanzoni, A. Stutt, and F. Ciravegna, "MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup," *Proc, 13th Int'l Conf. on Knowledge Engineering and Management*, pp. 379-391, 2002.

[19] S. Handschuh, S. Staab, and F. Ciravogna, "S-CREAM-- Semi-automatic CREAtion of Metadata in SAAKM 2002," *proc, Semantic Authoring, Annotation & Knowledge Markup Workshop*, pp. 27-33, 2002.

[20] A. Dingli, F. Ciravegna, and Y. Wilks, "Automatic Semantic Annotation using Unsupervised Information Extraction and Integration," *Proc, K-CAP 2003 Workshop on Knowledge Markup and Semantic Annotation*, pp. 2-9, 2003.

[21] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov, "KIM – Semantic Annotation Platform," *Proc, 2nd Int'l Semantic Web Conf.*, pp. 834-849, 2003.

[22] M. Ruiz-Casado, E. Alfonseca, and P. Castells, "From Wikipedia to Semantic Relationships: a Semi-automated Annotation Approach," *Proc, 1st Workshop on Semantic Wikis: From Wiki to Semantics, at the 3rd European Semantic Web Conf.*, Vol. 206 of Workshop on Semantic Wikis, 2006.

[23] M. Ruiz-Casado, E. Alfonseca, M. Okumura, and P. Castells, "Information Extraction and semantic annotation of Wikipedia," *Ontology Learning and Population: bridging the gap between text and knowledge*, pp. 145- 169, 2008.

[24] C. Jonquet, N. H. Shah, and M. A. Musen, "The Open Biomedical Annotator," *Proc, AMIA Summit on Translational Bioinformatics*, pp. 56-60, 2009.

[25] م. شمس‌فرد، *طراحی مدل یادگیر هستان‌شناسی: نمونه‌سازی در یک سیستم درک متن فارسی*، رساله دکتری، دانشگاه صنعتی امیرکبیر، ۱۳۸۱.

[26] T. Riddle, Parse::mediawikidump, URL: http://search.cpan.org/~triddle/Parse-MediaWikiDump-0.40/, 2006.

[27] T. Zesch, I. Gurevych, and M. Mühlhäuser, "Analyzing and Accessing Wikipedia as a Lexical Semantic Resource," *Proc, Biannual Conf. of the Society for Computational Linguistics and Language Technology*, pp. 213-221, 2007.

[28] F. Suchanek, G. Kasneci, and G. Weikum, "YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia," *Proc, 16th Int'l Conf. on World Wide Web*, pp. 697-706, 2007.

[4] Ch. Patel, K. Supekar, and Y. Lee, "OntoGenie: Extracting Ontology Instances from WWW," *Human Language Technology for the Semantic Web and Web Services*, pp. 123-126, 2003.

[5] V. De Boer, M. Van Someren, and B. J. Wielinga, "Relation instantiation for ontology population using the web," *Lecture Notes in Artificial intelligence (LNAI)*, Vol. 4314, pp. 202-213, 2007.

[6] P. Cimiano, and S. Staab, *Learning by Googling*, ACM press, 2004.

[7] A. Hearst, "Automatic Acquisition of Hyponyms from Large Text Corpora," *Proc, 14th Int'l Conf. on Computational Linguistics*, pp. 539-545, 1992.

[8] G. Geleijnse, and J. Korst, "Automatic Ontology Population By Googling," *Proc, 7th Belgium-Netherlands Conf. on Artificial Intelligence*, pp. 120-126, 2005.

[9] R. Bunescu, and M. Pasca. "Using encyclopedic knowledge for named entity disambiguation," *Proc, the European Conf. of the Association for Computational Linguistics*, pp. 9-16, 2006.

[10] S. Chernov, T. Iofciu, W. Nejdl, and X. Zhou, "Extracting Semantic Relationships between Wikipedia Categories," *Proc, 1st Int'l Workshop: SemWiki2006 — From Wiki to Semantics*, Vol. 206 of Workshop on Semantic Wikis, 2006.

[11] S. Soltani, and A. Abdollahzadeh Barforoush, "OILSW: A New System for Ontology Instance Learning in Semantic Web," *Proc, Int'l Conf. on Semantic Web and Digital Libraries*, pp. 54-63, 2007.

[12] P. Cimiano, A. Pivk, L.S. Thieme, and S. Staab, "Learning Taxonomic Relations from Heterogeneous Sources of Evidence," *Workshops on Ontology Learning from Text: Methods, Evaluation and Applications*, Vol. 123, pp. 59-73, 2004.

[13] P. Velardi, R. Navigli, A. Cuchiarelli, and F. Neri. "Evaluation of Ontolearn, a Methodology for Automatic Population of Domain Ontologies," *Workshops on Ontology Learning from Text: Methods, Evaluation and Applications*, pp. 92-106, 2005.

[14] H. Tanev, and B. Magnini, "Weakly Supervised Approaches for Ontology Population," *11th Conf. the European Chapter of the Association for Computational Linguistics, pp.* 17-24, 2006.

[15] D. Celjuska, and M. Vargas-Vera, "*Ontosophie: A Semi-Automatic System for Ontology Population from Text*," *Proc, Int'l Conf. on Natural Language Processing*, 2004.

[16] B. Magnini, E. Pianta, O. Popescu, and M. Speranza, "Ontology Population from Textual Mentions: Task Definition and Benchmark," *Proc, Workshop On Ontology*

مهرنوش شمس‌فرد دانش آموخته‌ی کارشناسی و کارشناسی‌ارشد رشته‌ی مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه صنعتی شریف و دکتری در رشته‌ی مهندسی کامپیوتر گرایش هوش مصنوعی از دانشگاه صنعتی امیرکبیر است. وی از سال ۱۳۸۲ با عنوان استادیار و سرپرست آزمایشگاه پردازش زبان طبیعی در دانشگاه شهید بهشتی مشغول به فعالیت‌های آموزشی - پژوهشی در زمینه‌های پردازش زبان طبیعی، مهندسی هستان‌شناسی، کاوش متن و وب معنایی است.
آدرس پست‌الکترونیکی ایشان عبارت است از:
m-shams@sbu.ac.ir

بهاره صراف‌زاده مدرک کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه شهید بهشتی تهران در سال ۸۷ دریافت نمود. او در حال حاضر در حال گذراندن دوره کارشناسی‌ارشد در دانشگاه یورک (تورنتو، انتاریو) می‌باشد و بر روی ترجمه ماشینی برای زبان فارسی - انگلیسی کار می‌کند. وی عضو آزمایشگاه پردازش زبان طبیعی دانشگاه شهید بهشتی است و علاقه اصلی او رشته‌ی هوش مصنوعی با تمرکز بر وب معنایی و درک متن می‌باشد.
آدرس پست‌الکترونیکی ایشان عبارت است از:
bahareh.sarrafzadeh@gmail.com

اطلاعات بررسی مقاله:

تاریخ ارسال: ۸۷/۰۶/۲۹
تاریخ اصلاح: ۸۹/۰۳/۰۹
تاریخ قبول شدن: ۸۹/۰۳/۲۶
نویسنده مرتبط: دکتر مهرنوش شمس‌فرد، دانشکده مهندسی برق و کامپیوتر، دانشگاه شهید بهشتی، تهران، ایران.

[1] Cross-Lingual
[2] Parallel Ontology Population and Text Annotation
[3] Ontology Population (or Instantiation)
[4] Lexemes
[5] lexicalize
[6] Cross-Language
[7] Google
[8] Wikipedia
[9] Hearst
[10] Head-Matching
[11] WordNet
[12] Semantic Annotation Platform
[13] Wrapper Induction
[14] Lexico-Syntactic
[15] Segmenter
[16] POS tags
[17] Token
[18] Proper Noun
[19] Common Noun
[20] Morphological
[21] Semantic Tags
[22] Head-Matching
[23] Stop Word
[24] Googling
[25] Precision
[26] Recall
[27] Application Program Interface
[28] Word Sense Disambiguation
[29] Synsets
[30] Sense