بیست و یکمین کنفرانس ملی سالانه انجمن کامپیوتر ایران
پژوهشکده علوم کامپیوتر، پژوهشگاه دانشهای بنیادی، فرمانیه، تهران ۱۸-۲۰ اسفند ماه ۱۳۹۴

ارائه روشی جهت بهبود دقت سامانه‌های استخراج آزاد اطلاعات با کمک ویژگی‌های رابطه در دامنه

وحیده رشادت[۱]، مریم حورعلی[۲]، هشام فیلی[۳]

[۱] پژوهشکده فناوری اطلاعات، دانشگاه صنعتی مالک اشتر، تهران، ایران
vreshadat@mut.ac.ir

[۲] استادیار، پژوهشکده فناوری اطلاعات، دانشگاه صنعتی مالک اشتر، تهران، ایران
mhourali@mut.ac.ir

[۳] استادیار، دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران، تهران، ایران
hfaili@ut.ac.ir

چکیده

استخراج آزاد اطلاعات روش استخراج مستقل از رابطه است که روابط را بطور مستقیم از مجموعه داده‌های بزرگ و غیرهمگن مانند وب استخراج می‌کند. برخلاف روش‌های پیشین برای استخراج اطلاعات، روش‌های استخراج آزاد اطلاعات نیاز به واژگان خاص یا دامنه‌های از قبل مشخص شده برای عمل استخراج ندارند و استخراج روابط دلخواه از جملات را در متن ممکن می‌سازند.

یک چالش اصلی برای سامانه‌های استخراج آزاد اطلاعات، تخمین احتمال درست بودن رابطه‌ی استخراج شده است. به دلایل متعددی از جمله افزایش کارایی الگوریتم‌های داده کاوی، بهبود یکپارچگی داده‌ها و استخراج اطلاعات محاوره‌ای، نیاز به معیار ضریب اطمینانی وجود دارد که نشان دهد رابطه‌ی استخراج‌شده نمونه‌ی درستی از رابطه‌ی بین موجودیت‌ها است.

در این مقاله تلفیقی از چندین ویژگی پیشنهادی مختلف برای انتساب معیار ضریب اطمینان با استفاده از رگرسیون منطقی دوجمله‌ای نشان داده شده است. این ویژگی‌ها برخی خواص استخراج نظیر تعداد اسنادی که رابطه از آنها استخراج شده، تعداد آرگومان‌های رابطه و نوع آنها را در نظر می‌گیرد. معیار ضریب اطمینان پیشنهادی به خروجی چندین سامانه استخراج آزاد اطلاعات اعمال شده و دقت نتایج تحت تاثیر ضریب اطمینان پیشنهادی، بررسی شده است. ارزیابی‌ها نشان می‌دهد که تلفیق ویژگی‌های مطرح‌شده امیدبخش است و دقت خروجی‌ها با اعمال روش پیشنهادی بالاتر از دقت حالت پایه است. بالاترین افزایش دقت مربوط به سامانه-های ReVerb و TextRunner است که افزایشی در حدود ۴٪دارد.

کلمات کلیدی

پردازش زبان طبیعی، استخراج اطلاعات، استخراج آزاد اطلاعات، استخراج رابطه، ضریب اطمینان

۱-مقدمه

استخراج اطلاعات فرایند استخراج خودکار داده‌های ساخت یافته از متن غیرساخت‌یافته است. یکی از وظایف اصلی در استخراج اطلاعات، استخراج رابطه است که هدف آن استخراج روابط معنایی بین موجودیت‌ها از متون زبان طبیعی است. نیاز به استخراج روابط نه تنها از حیاتی‌ترین موارد در فهم معنای متن برای ماشین‌هاست بلکه می‌تواند در کاربردهای زیادی مانند جستجوی وب، پرسش‌وپاسخ، داده‌کاوی، ساخت پایگاه دانش، ساخت هستان‌نگار[۱] درک نیت نویسنده متن، اخبار (شیوع بیماری، حملات تروریستی و سایر اطلاعات)، پزشکی نیز بکار رود. اطلاعات غیرساخت‌یافته قابل خواندن، سازماندهی و تحلیل توسط ماشین‌ها نیستند. بدلیل افزایش روزافزون حجم زیاد اطلاعات در وب جهان‌گستر که اغلب به شکل متن غیرساخت‌یافته ذخیره شده‌اند این مشکل تشدید شده و استخراج خودکار روابط مورد توجه زیادی قرار گرفته است. استخراج روابط، اصلی‌ترین بخش استخراج اطلاعات به شمار می‌رود. در این وظیفه روابط معنایی بین موجودیت‌ها کشف می‌شود [۱, ۲].

مقیاس بزرگ و در حال رشد، ترکیب گونه‌های مختلف از اسناد و انواع نامحدودی از روابط از جمله چالش‌های استخراج روابط در مقیاس وب است [۳]. روش‌های سنتی برای استخراج اطلاعات فرض می‌کنند که مجموعه‌ی ثابتی از روابط موردنظر از قبل مشخص شده‌اند. این روش‌ها معمولا قابل گسترش به مقیاس وب که در آن تعداد روابط موردنظر بسیار بزرگ است نیستند [۴]. یک روش جایگزین استخراج آزاد اطلاعات است که هدفش این است که روش‌های استخراج اطلاعات را از جهت اندازه و تنوع به مقیاس وب سوق دهد. استخراج آزاد اطلاعات از استخراج اسم‌ها و افعال خاص و از پیش تعریف‌شده جلوگیری می‌کند و استخراجگرها در این سیستم‌ها غیرلغوی هستند. این روش‌ها اغلب خود ناظر[۲] هستند و با ایجاد خودکار دادگان آموزشی با استفاده از دسته‌بند و به کمک ویژگی‌های مختلف، روابط را تشخیص می‌دهند [۵]. اهداف کلیدی در استخراج آزاد اطلاعات عبارتنداز: (۱) مستقل از دامنه‌بودن (۲) استخراج بدون ناظر (۳) مقیاس‌پذیر بودن به حجم زیادی از متون [۶].

از آنجایی که استخراج آزاد اطلاعات هرگز بطور کامل دقیق نیست، داشتن معیار ضریب اطمینان موثر، مفید به نظر می‌رسد. مطابق [۷] حداقل سه کاربرد مهم برای تخمین ضریب اطمینان وجود دارد. اول، مسامحه بین دقت و پوشش یک روش عادی برای بهبود یکپارچگی داده در پایگاه‌های داده است. ایجاد موثر این مسامحه نیاز به پیش‌بینی دقیق معیار درستی دارد. دوم، تخمین‌های ضریب اطمینان برای استخراج اطلاعات محاوره‌ای ضروری است که در آن ممکن است کاربران فیلدهای نادرست استخراج‌شده را تصحیح کنند. این اصلاحات سپس بطور خودکار به منظور تصحیح دیگر خطاها در همان رکورد انتشار می‌یابند. هدایت کاربر به فیلدی با حداقل ضریب اطمینان، به سامانه اجازه می‌دهد تا کارایی‌اش را با حداقل تلاش کاربر بهبود بخشد. سوم، تخمین ضریب اطمینان می‌تواند کارایی الگوریتم‌های داده کاوی را بهبود بخشد که به پایگاه‌داده‌هایی که توسط سامانه‌های استخراج اطلاعات ایجاد می‌شوند، بستگی دارد. تخمین ضریب اطمینان برای برنامه‌های کاوش داده مجموعه‌ی قویتری از فرض‌های پایین به بالا[۳] را فراهم می‌کند و باعث می‌شود تا استنباط‌های دقیق‌تری بوجود آید.

این مقاله به ارائه‌ی روشی جهت تخمین ضریب اطمینان در سامانه‌های استخراج آزاد اطلاعات تمرکز دارد. در این کار از رگرسیون منطقی[۴]، که یک مدل یادگیری ماشینی احتمالی است برای انتساب خودکار وزن ضریب اطمینان به یک استخراج استفاده شده است. این مدل می‌تواند ویژگی‌های پیشنهادی را بگیرد و احتمال اینکه مشاهده‌ی خاصی در کلاس صحیح است را برگرداند. این مقاله نوآوری‌های زیر را دارد:

- چندین ویژگی مهم پیشنهاد شده است که تعدادی از خواص رابطه‌ی استخراج‌شده از جمله تعداد اسناد مجزایی که استخراج از آنها گرفته می‌شود، تعداد آرگومان‌های رابطه، نوع آنها را در نظر می‌گیرد.
- اینکه چگونه ویژگی‌های پیشنهادی برای وزندهی روابط استخراج‌شده، دقت نتایج را تحت تاثیر قرار می‌دهند، مورد مطالعه قرار گرفته است و از یک دسته‌بند رگرسیون منطقی که روی داده‌های نمونه با ویژگی‌های پیشنهادی، آموزش داده‌شده است برای انتساب ضریب اطمینان به هر استخراج صورت گرفته توسط استخراج آزاد اطلاعات، استفاده شده است تا دقت را بهبود بخشد.
- نتایج آزمایش‌های صورت گرفته بر روی خروجی سه سامانه‌ی استخراج آزاد اطلاعات نشان می‌دهد که روش پیشنهادی می‌تواند خروجی‌های نوفه‌دار[۵] را از خروجی سامانه‌ها کاهش داده و در نتیجه باعث بهبود دقت شود.

ساختار مقاله در این قالب است. در بخش دوم کارهای پیشین انجام شده، بیان شده است. در بخش سوم روش پیشنهادی بطور کامل شرح داده شده است. در بخش چهارم آزمایش‌های صورت گرفته بر روی خروجی سه سامانه‌ی استخراج آزاد اطلاعات نشان شده است و در بخش پایانی، نتیجه‌گیری لازم ارائه شده است.

۲- پیش‌زمینه وکارهای مرتبط

در این بخش ابتدا توضیح مختصری از استخراج آزاد اطلاعات و سامانه‌های به‌کار رفته در آزمایش‌ها ارائه خواهد شد و سپس کارهای پیشین و مرتبط بیان خواهد شد.

استخراج آزاد اطلاعات یک پارادایم جدید در استخراج اطلاعات است که به مجموعه‌ی کوچک و شناخته شده‌ای از روابط هدف محدود نیست. این وظیفه یکی از موفقیت‌های خواندن ماشینی است. سیستم‌های استخراج آزاد اطلاعات به موفقیت قابل توجهی روی پیکره‌های بزرگ و دامنه باز مانند وب دست یافته‌اند. و یک چالش مهم برای سامانه‌های استخراج آزاد اطلاعات مشخص کردن احتمال درست بودن اطلاعات استخراج‌شده است. در این مقاله از یک روش یادگیری ماشینی به منظور انتساب ضریب اطمینان برای خروجی سامانه‌های استخراج آزاد اطلاعات استفاده شده است. در ادامه سه سامانه‌ی بکار رفته در بخش آزمایش‌ها به اختصار شرح داده خواهند شد.

---

۱ Ontology

۲ self-supervised

۳ bottom-up

۴ logistic regression

۵ noisy

TextRunner [۸]: از اولین سامانه‌های استخراج آزاد اطلاعات بوده است که می‌تواند تعداد نامحدود روابط را با یک گذر در مقیاس وب استخراج کند. این سیستم مستقل از دامنه است و یک رابطه و آرگومان‌های آن را با روش خودناظر استخراج می‌کند. در واقع این سامانه از داده‌هایی که خودش برچسب‌زده است، استفاده می‌کند، تا عبارت‌های رابطه‌ای را بیابد و یک مدل از نوع دسته‌بند که مشخص کننده‌ی وجود یا عدم وجود رابطه است، تولید می‌کند. در این روش، دادگان آموزشی با ویژگی‌های عمیق و دسته‌بند با ویژگی‌های سطحی ایجاد شده است.

ReVerb [۹]: از سریع‌ترین و موفق‌ترین سامانه‌های استخراج آزاد اطلاعات است که سه ویژگی مهم دارد. ۱) در استخراج نام رابطه، با در نظر گرفتن کل کلمات جمله، رابطه با استفاده از قیدهای واژگانی و نحوی استخراج می‌شود. این قواعده نحوی از ویژگی‌های برچسب اجزای کلام بهره می‌گیرند. ۲) از یک واژه‌نامه‌ی روابط استفاده می‌شود، تا روابط خیلی خاص استخراج نشوند. ۳) به جای این که ابتدا آرگومان ها استخراج شوند، ابتدا نام رابطه استخراج می‌شود و سپس آرگومان‌های آن استخراج می‌شوند.

WOE_{POS} [۱۰]: از روش خاصی برای آموزش استخراجگر که اصطلاحاً نظارت دور گفته می‌شود، استفاده می‌کند. در این سیستم از اطلاعات موجود در جعبه‌های اطلاع[۶] ویکی‌پدیا استفاده می‌شود. هر اطلاع یک رابطه‌ی دوتایی است که یکی از آرگومان‌های آن موضوع صفحه‌ی ویکی پدیا و دیگری مقادیر صفات آن است. با انطباق اطلاعات با جملات متن، جملات و رابطه‌ی استخراج شده از آن‌ها به دست می‌آید و به عنوان داده آموزشی مورد استفاده قرار می‌گیرد. در واقع WOE_{POS} مثال‌های آموزشی خاص–رابطه را با تطبیق مقادیر صفات جعبه‌های اطلاع با جملات مربوطه تولید می‌کند اما WOE_{POS} این نمونه‌ها را به دادگان آموزشی مستقل از رابطه تبدیل می‌کند تا استخراجگر غیرلغوی (مستقل از لغت) یادگیری شود. WOE_{POS} فقط محدود به ویژگی‌های سطحی مانند برچسب‌گذاری اجزای کلام بوده و همانند TextRunner سریع است.

بیشتر کارهای انجام شده برای تخمین ضریب اطمینان برای استخراج اطلاعات از روش‌های یادگیری استفاده کرده‌اند. آقای شفر و همکارانش [۱۱] با کمک مدل‌های مارکوف پنهان برای تخمین ضریب اطمینان در استخراج اطلاعات استفاده کردند. آنها ضریب اطمینان را برای همه‌ی فیلدها تخمین نزدند بلکه تنها برای هر توکن‌های یگانه این کار انجام شد. آنها اطمینان یک توکن را توسط اختلاف بین احتمال اولین و دومین برچسب‌های محتمل آن تخمین زدند. روش‌های استخراج مبتنی بر قاعده، اطمینان را براساس پوشش قاعده در دادگان آموزش تخمین می‌زنند. دیگر زمینه‌هایی که در آنها تخمین اطمینان بکار می‌رود شامل دسته‌بندی اسناد [۱۲]، که دسته‌بندها با استفاده از ویژگی‌های اسناد ساخته می‌شوند. تشخیص گفتار [۱۳] که اطمینان برای کلمه‌ی تشخیص داده شده توسط لیستی از کلماتی تخمین زده می‌شود که معمولا در تشخیص آنها اشتباه رخ می‌دهد. در ترجمه‌ی ماشینی نیز از روش-های مختلفی از جمله از شبکه‌های عصبی برای یادگیری احتمال ترجمه‌ی صحیح یک کلمه با استفاده از ویژگی‌های متن استفاده می‌شود.

در [۷] از مدل میدان‌های تصادفی شرطی[۷] استفاده شده است که نوعی مدل گرافیکی است که بطور خودکار فیلدهای رکوردها را برچسب می‌زند. در اینجا رکورد، یک بلاک کامل از اطلاعات شخص و فیلد یک جزئی از آن رکورد است. از چندین روش برای تخمین ضریب اطمینان فیلد و رکورد استفاده شده است. در این روش از یک تخمین‌زن اطمینان مبتنی بر ریاضی برای سامانه‌های استخراج اطلاعات با وضعیت متناهی استفاده شده است. سامانه استخراج اطلاعاتی بررسی شده بر اساس مدل میدان‌های تصادفی شرطیِ خطی–زنجیری است و نتایج حاصل، بهبود قابل توجهی در دقت تخمین درستی فیلد نشان می‌دهد.

در [۱۴] یک مدل ترکیبی بنام URNS پیشنهاد شده است که تاثیر اندازه‌ی نمونه، فراوانی و اعتبارسنجی از چندین قاعده‌ی استخراج مجزا، روی احتمال صحیح بودن یک استخراج مورد بررسی قرار گرفته است. این روش از داده‌های برچسب‌زده‌شده‌ی دستی استفاده نمی‌کند. نتایج آزمایش‌ها نشان می‌دهد که این روش نسبت به روش‌های بدون ناظر بهتر عمل می‌کند.

در TextRunner [۸] که اولین سامانه‌ی استخراج آزاد اطلاعات است و کارایی بالایی در مدیریت حجم عظیم اطلاعات دارد از یک ارزیاب استفاده شده است. این ارزیاب تعداد جملات مجزایی را در نظر می‌گیرد که از آنها استخراجی یافت می‌شود. ارزیاب با کمک مدل احتمالاتی که قبلا در سامانه استخراج اطلاعات بدون ناظر KNOWITALL [۱۵] بکار رفته است از شمارش این جملات برای انتساب احتمال به هر سه‌تایی استخراجی استفاده می‌کند. این روش در مقابل روش‌های دیگری نظیر [۱۶] که بر مبنای میزان باهم‌آیی[۸] است بهتر عمل می‌کند.

KNOWITALL [۱۵] یک سامانه استخراج اطلاعات بدون ناظر و مستقل از دامنه است که هدفش پردازش خودکار حجم بزرگی از اطلاعات در مقیاس وب و استخراج حقایق(برای مثال اسامی دانشمندان یا سیاستمداران) است. توانایی استخراج اطلاعات بدون نمونه‌های آموزشی که بصورت دستی برچسب‌زده شده است آن را از سامانه‌های قبلی جدا ساخته است. این سامانه از این فرضیه استفاده می‌کند که استخراج‌هایی که از جملات مجزای فراوانتری در پیکره گرفته می‌شوند احتمال درست بودن‌شان زیاد است.

خروجی‌های روش استخراج رابطه‌ی نیمه ناظر نوفه‌دار است و نیاز به تخمین کیفیت اطلاعات استخراج‌شده امری ضروری است. در [۱۷] روشی برای بهبود پیشنهاد شده است که از الگوریتم بیشینه‌سازی امید ریاضی[۹] برای ارزیابی خودکار کیفیت الگوهای استخراجی و سه‌تایی‌های رابطه‌ی بدست آمده استفاده شده است. موثر بودن این روش روی گستره‌ی وسیعی از روابط بررسی شده است.

همانطوری که قبلا ذکر شد Reverb [۹] بعنوان یکی از بهترین و مقاوم‌ترین استخراجگرهای آزاد اطلاعات تا کنون است که از دنباله‌ای از برچسب‌های اجزای کلام بعنوان یک محدودیت نحوی برای استخرج عبارات رابطه‌ای، حذف استخراج‌های غیر منسجم[۱۰] و حاوی اطلاعات بی‌معنی[۱۱]

---

[۶] infobox

[۷] Conditional Random Field

[۸] Pointwise Mutual Information

[۹] Expectation Maximization

[۱۰] incoherent extractions

[۱۱] uninformative extractions

استفاده می‌کند. این سامانه از یک تابع اطمینان با کمک یک دسته‌بند برای انتساب امتیاز اطمینان استفاده می‌کند. این دسته‌بند از تعدادی از ویژگی‌های مستقل از رابطه شامل تعداد کلمات، اسم خاص بودن یا نبودن، بررسی نوع حروف اضافه استفاده کرده است.

۳-روش پیشنهادی

در این بخش روش پیشنهادی ما برای انتساب احتمال درستی به استخراج-های سامانه‌های استخراج آزاد اطلاعات شرح داده خواهد شد. پارامترهای مختلفی وجود دارد که می‌تواند در تشخیص روابطِ دقیق کمک کند. بر این اساس یک روش مبتنی بر یادگیری ارائه شده است که از پارامترهای پیشنهادی بعنوان ویژگی استفاده می‌کند تا وزنی را برحسب درستی به روابط معنایی استخراج‌شده انتساب کند. بر اساس این فرض، استخراج‌ها با دقت بالا بدست خواهند آمد. در شکل (۱) مولفه‌های اصلی روش پیشنهاد شده نشان داده شده است.

خروجی سامانه‌های استخراج آزاد اطلاعات که بصورت روابط است بعنوان ورودی روش پیشنهادی است. در قسمت دسته‌بند از یک رگرسیون منطقی استفاده شده است. رگرسیون منطقی یک مدل احتمالاتی شرطی است که برای یک نمونه داده شده، دسته‌بند احتمالی توزیع احتمالی روی همه کلاس-ها را تولید می‌کند [۱۸].

دسته‌بند احتمالاتی رگرسیون منطقی دوجمله‌ای روی یک مجموعه‌ی برچسب‌زده شده با ویژگی‌های پیشنهادی آموزش داده می‌شود و از وزن دسته‌بند برای کلاس صحیح برای انتساب یک امتیاز اطمینان به هر رابطه استخراج شده استفاده می‌شود. چندین ویژگی برای این دسته‌بند در نظر گرفته شده است که در ادامه توضیح داده خواهد شد.

فراوانی سند: این ویژگی بر اساس این فرض است که یک رابطه‌ی معتبر به طور مکرر در اسناد مختلفی در مقیاس بزرگ مانند وب دیده می‌شود. به خصوص، این ویژگی تاثیر افزونگی روی احتمال درستی را در نظر می‌گیرد و به عنوان تعداد اسناد مجزایی که از آنها هر استخراج پیدا می‌شود نسبت به تعداد کل اسناد تعریف می‌شود.

$$F_1 = \frac{|D_r|}{|D|} \qquad (۱)$$

$|D|$ تعداد کل اسناد و $|D_r|$ تعداد اسنادی که شامل رابطه‌ی r است. هر چه رابطه‌ی استخراج شده از تعداد اسناد بیشتری استخراج شود، مقدار F_1 بیشتری می‌گیرد.

فراوانی نوع: این ویژگی تعداد دامنه‌هایی که در آنها رابطه ظاهر می‌شود را در نظر می‌گیرد. فراوانی نوعِ آرگومان‌های رابطه که رابطه در متن آنها ظاهر می-شود، بکار گرفته می‌شود. این ویژگی بصورت زیر نشان داده شده و تعریف می‌شود.

$$F_2 = \frac{|T_r|}{|T|} \qquad (۲)$$

$|T_r|$ تعداد نوعِ آرگومان‌های مجزایی که یک رابطه می‌گیرد و در واقع رابطه از آنها استخراج شده است و $|T|$ تعداد کل نوع‌های آرگومان‌ها است.

این ویژگی عام و خاص بودن رابطه در دامنه را درنظر می‌گیرد و هر چه رابطه‌ی استخراج شده با نوع آرگومان‌های بیشتری همراه شود، مقدار F_2 بیشتر است. برای محاسبه‌ی این ویژگی از تشخیص موجودیت‌های اسمی مربوط به بسته‌ی OpenNLP[۱۲] استفاده شده است که شامل ۸ نوع مختلف (مانند زمان، مکان، شخص) است.

فراوانی آرگومان: بر طبق Reverb [۱۹] هرچه رابطه با تعداد بیشتری جفت مختلف آمده باشد احتمال رابطه بودن آن زیاد است. این معیار نیز بر حسب کل جفت آرگومان‌ها بدست آمده است. این ویژگی براساس تعداد آرگومان-های مجزایی است که یک رابطه می‌گیرد و بصورت زیر تعریف می‌شود:

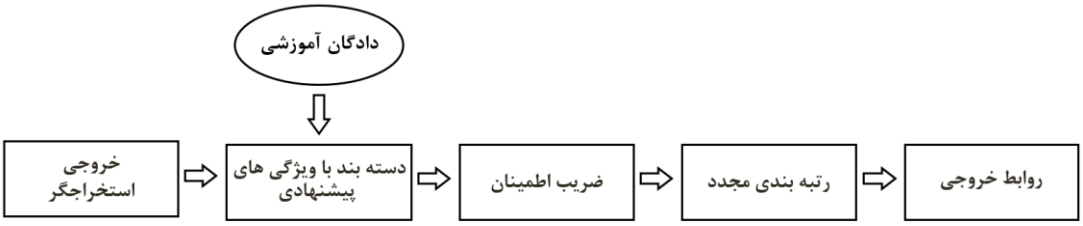
$$F_3 = \frac{|A_r|}{|A|} \qquad (۳)$$

$|A|$ تعداد کل آرگومان‌های مجزا در مجموعه متونی که از آنها استخراج صورت گرفته و $|A_r|$ تعداد آرگومان‌های مجزایی است که رابطه می‌گیرد.

برخلاف روش‌های پیشینِ محاسبه‌ی ضریب اطمینان، در معیار پیشنهادی، نوع آرگومان‌های رابطه (دامنه‌ی رابطه) و تکرار رابطه در دامنه‌ی موردنظر نیز مورد توجه قرار می‌گیرد. ویژگی‌های استفاده شده در دسته‌بند، بصورت کارا قابل محاسبه است. در این مقاله چگونگی استفاده از این ویژگی‌ها برای وزندهی روابط استخراج‌شده و تحت تاثیر قرار گرفتن دقتِ نتایج بررسی شده است. در بخش بعدی جزئیات بیشتری درباره‌ی نتایج آزمایش‌ها ارائه خواهد شد.

۴- آزمایش‌ها و ارزیابی روش

تاثیر بکارگیری دسته‌بند رگرسیون منطقی، یک روش رگرسیون غیرخطی، روی خروجی سامانه‌های ReVerb و WOE_{POS} و نیز TextRunner با کمک ویژگی‌های پیشنهادی ارزیابی شده و رفتار آن بررسی و مقایسه شده است.



شکل (۱) : معماری کلی مولفه‌ی ارزیاب رابطه

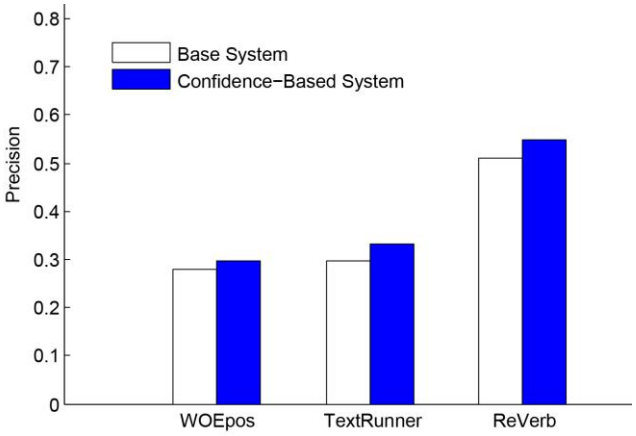
در این قسمت از مجموعه داده‌ای استفاده شده است که توسط آقای فادر و همکارانش [۱۹] تهیه شده است. آنها یک مجموعه داده شامل ۵۰۰ جمله از

[۱۲] https://opennlp.apache.org/

وب با کمک سرویس لینک تصادفی یاهو[۱۳] ایجاد کردند. این مجموعه‌ی داده شامل خروجی استخراجگرهای مختلف (نظیر TextRunner و ReVerb) روی ۵۰۰ جمله انتخابی است. دو داور انسانی بطور دستی و مستقل هر استخراج را به طور مستقل ارزیابی کردند و برچسبی بصورت «درست» یا «نادرست» زدند. این داورها روی ۸۶٪ از استخراج‌ها با امتیاز توافقی در حدود κ=۶۸٪ به توافق رسیدند. زیرمجموعه‌ای از داده که دو داور به توافق رسیده‌اند، مورد استفاده قرار گرفته است. داورها استخراج‌هایی را که اطلاعی در بر نداشتند (آنهایی که اطلاعات اساسی از آنها حذف شده بودند) بعنوان استخراج نادرست در نظر گرفتند. این روشِ برچسب‌زدن سخت‌گیرانه‌تر از آنچه است که قبلا در برچسب‌زنی در ارزیابی‌های سامانه‌های استخراج اطلاعات بکار گرفته شده است [۱۹].

در این مجموعه، استخراج‌ها از مجموعه‌ای از ۱۰۰۰ جمله از وب و ویکی‌پدیا نیز بطور دستی بصورت درست یا نادرست برچسب زده شده‌اند. دسته‌بند روی ۱۰۰۰جمله تصادفی با ویژگی‌های پیشنهادی آموزش داده شده است. از آنجا که مجموعه‌ی داده شامل جملات است، فراوانی اسناد با این فرض انجام شده است که هر کدام از جملات به عنوان یک سند مجزا در نظر گرفته شوند. در اینجا از بسته‌ی OpenNLP برای برچسب‌زنی موجودیت‌های اسمی آرگومان‌ها استفاده شده است و تمام پیاده‌سازی‌ها نیز در محیط جاوا انجام گرفته است.

با آموزش دسته‌بند رگرسیون منطقی دوجمله‌ای با ویژگی‌های گفته شده روی مجموعه‌ی دادگان ذکر شده، یک امتیاز اطمینان به هر سه‌تایی انتساب می‌شود. احتمال تعلق به کلاس صحیح بعنوان این امتیاز در نظر گرفته می‌شود. TextRunner، ReVerb و WOE$_{PARSE}$ روی مجموعه‌ی تست ۵۰۰ جمله‌ای اجرا و نتایج استخراج بررسی شده است. مقادیر مختلف حدآستانه ضریب اطمینان را برای ارزیابی تغییرات دقت بکار بردیم.



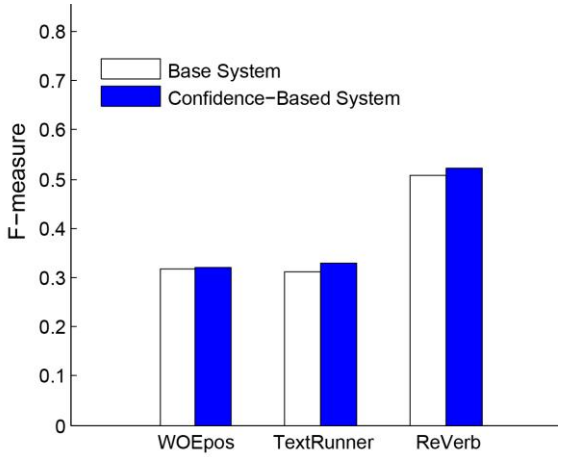
شکل (۲) : مقادیر دقت در بکارگیری روش پیشنهادی در خروجی سامانه‌های استخراج آزاد اطلاعات در مقایسه با حالت پایه

دقت بصورت نرخ تعداد استخراج‌های درست بازیابی‌شده به تعداد کل استخراج‌های بازیابی شده تعریف می‌شود. نتایج اولیه از تحلیل دقت در شکل (۲) گزارش شده است.

[۱۳] http://random.yahoo.com/bin/ryl

تعداد استخراج‌های با امتیاز بالا با افزایش امتیاز اطمینان کاهش می‌یابد. با افزایش مقادیر حد آستانه ضریب اطمینان افزایش می‌یابد و تعداد استخراج‌های درست افزایش می‌یابد تاجایی که تقریبا تمام استخراج‌های انجام‌شده در حدآستانه‌های بالا درست هستند. با آزمایش‌های صورت گرفته بهترین حدآستانه ۰٫۸ است. در این حالت دقت خروجی نسبت به حالت پایه برای هر سامانه افزایش پیدا کرده است. این امر موثر بودن ویژگی‌های پیشنهادی را نشان می‌دهد. بیشترین افزایش مربوط به سامانه ReVerb و TextRunner است که دقت نسبت به حالت پایه حدود ۴٪ افزایش پیدا کرده است. نتایج آزمایش‌ها نشان می‌دهد که روش پیشنهادی می‌تواند خروجی‌های نوفه‌دار را از خروجی کاهش داده و در نتیجه باعث بهبود دقت شود. بنظر می‌رسد با افزایش تعداد ویژگی‌های موثر و اندازه‌ی دادگان آموزش بتوان بهبود بیشتری در نتایج حاصل داد.

با افزایش مقادیر حدآستانه، بازخوانی نیز به آرامی کاهش می‌یابد. در ادامه برای بررسی کارایی سیستم، مقدار امتیاز-f نیز بررسی شده است. امتیاز-f تلاشی برای یافتن مسامحه بین دقت و بازخوانی است. در شکل (۳) نتایج تحلیل مشخص شده است.



شکل (۳) : مقادیر امتیاز-f در بکارگیری روش پیشنهادی در خروجی سامانه‌های استخراج آزاد اطلاعات در مقایسه با حالت پایه

همانطور که مشاهده می‌شود، بجز سامانه WOEpos که افزایش امتیاز-f آن بسیار ناچیز است، TextRunner و ReVerb هر کدام افزایشی در حدود ۱٫۵٪ داشته‌اند. این موضوع نشان می‌دهد که اعمال روش پیشنهادی باعث افزایش کارایی سامانه‌ها شده و می‌تواند به عنوان روشی موثر در امتیازدهی خروجی‌ها بکار رود.

۵- نتیجه‌گیری و کارهای آینده

تقریبا تمام سامانه‌های استخراج آزاد اطلاعات دارای خطا هستند و یک چالش مهم برای سامانه‌های استخراج آزاد اطلاعات مشخص کردن احتمال درست بودن اطلاعات استخراج‌شده است.

در این مقاله از یک دسته‌بند رگرسیون منطقی دوجمله‌ای برای تخمین یک امتیاز اطمینان برای رابطه‌های استخراج شده توسط سامانه‌های استخراج آزاد اطلاعات استفاده شده است که ویژگی‌های مختلفی از رابطه استخراجی را در نظر می‌گیرد. ویژگی‌های پیشنهادی برخی از خواص رابطه از جمله تعداد

آرگومان‌های رابطه و نوع آنها را در نظر می‌گیرد. آزمایش‌ها نشان می‌دهد که ویژگی‌های پیشنهادی تعداد خروجی‌های نادرست را کاهش داده و در نتیجه باعث بهبود دقتِ نتایج حاصل می‌شود.

در آینده، بدنبال استفاده از ویژگی‌های بیشتر به منظور بهبود کارایی مدل یادگیری شده خواهیم بود. علاوه‌بر این در حال توسعه‌ی آزمایش‌ها برای سایر سامانه‌های استخراج اطلاعات و نیز بکارگیری روش پیشنهادی در کاربردهای دیگری نظیر داده کاوی، اعمال به خروجی آنها و بررسی نتایج حاصل خواهیم بود. از طرفی تاثیر اندازه دادگان آموزشی و انجام آزمایش‌ها با دادگان آزمایشی بزرگتر نیز بررسی خواهد شد. دیدگاه دیگر برای بهبود نتایج می‌تواند بسط فضای نوع با منابعی از دانش معنایی مانند هستان‌نگارها باشد.

مراجع

[1] Piskorski, J., Yangarber R., *Information extraction: Past, present and future*, Multi-source, Multilingual Information Extraction and Summarization. Springer, pp. 23-49, 2013.

[2] Yao L, Haghighi A., Riedel S. and McCallum A., *Structured relation discovery using generative models*, In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1456-146, 2012.

[3] Min B., Shi S., Grishman R. and Lin C-Y., "*Towards Large-Scale Unsupervised elation Extraction from the Web"*, International Journal on Semantic Web and Information Systems (IJSWIS), 8(3): 1-23, 2012.

[4] Banko M., Etzioni O. and Center T., "The Tradeoffs Between Open and Traditional Relation Extraction", In ACL, pp. 28-36, 2008.

[5] Schmitz M., Bart R., Soderland S. and Etzioni O. "*Open language learning for information extraction"*, In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 523-534., 2012.

[6] Del Corro L., Gemulla R., *"ClausIE: clause-based open information extraction"*, In Proceedings of the 22nd international conference on World Wide Web, pp. 355-366, 2013.

[7] Culotta A. and McCallum A., *"Confidence estimation for information extraction"*, In Proceedings of HLT-NAACL 2004: Short Papers, pp. 109-112, 2004.

[8] Banko M., Cafarella M J., Soderland S., Broadhead M. and Etzioni O., Open information extraction for the web. In IJCAI, pp. 2670-2676, 2008.

[9] Etzioni O., Fader A., Christensen J., Soderland S. and Mausam M. *"Open Information Extraction: The Second Generation"*, In IJCAI, pp. 3-10, 2011.

[10] Wu F. and Weld D S., *"Open information extraction using Wikipedia"*, In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 118-127, 2010.

[11] Scheffer T., Decomain C. and Wrobel S., *"Active hidden markov models for information extraction"*, Advances in Intelligent Data Analysis. Springer, pp. 309-318, 2001.

[12] Bennett P. N., Dumais S. T. and Horvitz E., *"Probabilistic combination of text classifiers using reliability indicators: Models and results"*, In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 207-214.

[13] Gunawardana A., Hon H-W. and Jiang L., *"Word-based acoustic confidence measures for large-vocabulary speech recognition"*, In ICSL, 1998.

[14] Downey D., Etzioni O. and Soderland S., *"Analysis of a probabilistic model of redundancy in unsupervised information extraction"*, *Artificial Intelligence*, 174(11): 726-748, 2010.

[15] Etzioni O., Cafarella M., Downey D., Popescu A-M., Shaked T., Soderland S., Weld D S. and Yates A., *"Unsupervised named-entity extraction from the web: An experimental study"*, *Artificial intelligence*, 165(1): 91-134, 2005.

[16] Downey D., Etzioni O. and Soderland S. *"A probabilistic model of redundancy in information extraction"*, 2006.

[17] Agichtein E., *"Confidence estimation methods for partially supervised relation extraction"*, In Proc. of SIAM Intl. Conf. on Data Mining (SDM06), 2006.

[18] Aggarwal C. C. and Zhai C., *Mining text data*, 2012.

[19] Fader A., Soderland S. and Etzioni O., "*Identifying relations for open information extraction*", In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1535-1545, 2011.