روشی مبتنی بر قاعده جهت بهبود کارایی سامانه‌های استخراج آزاد اطلاعات با استفاده از درخت تجزیه‌ی وابستگی

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

چکیده

استخراج آزاد اطلاعات بر خلاف روش‌های پیشینِ استخراج اطلاعات، از معماری‌هایی که نیاز به مشخص‌کردن روابط از قبل دارند جلوگیری می‌کنند و محدود به روابط خاصی نیستند. بنابراین قادر به استخراج روابط دلخواه بطور مستقیم از مجموعه داده‌های بزرگ و دامنه‌های غیرهمگن مانند وب هستند. یک چالش اصلی برای سامانه‌های استخراج آزاد اطلاعات این است که روابط استخراج‌شده نمونه‌های درستی از روابط بین موجودیت‌ها باشد. نیاز به استخراجگری که بتواند با اطمینان بالا به کشف اطلاعات بپردازد، از جمله اهداف استخراج آزاد اطلاعات است. در این راستا، در این مقاله روشی مبتنی بر قاعده پیشنهاد شده است که با کمک ویژگی‌های جمله و درخت تجزیه‌ی وابستگی، منجر به افزایش خروجی‌های درست و کاهش خروجی‌های نادرست و در نتیجه افزایش دقت و بازخوانی می‌شود. روش پیشنهادی به خروجی چندین سامانه استخراج آزاد اطلاعات اعمال شده و دقت نتایج تحت تاثیر آن بررسی شده است. ارزیابی‌ها نشان می‌دهد که روش پیشنهادی امیدبخش است و معیارهای کارایی خروجی‌ها با اعمال این روش بالاتر از حالت پایه است.

کلمات کلیدی

پردازش زبان طبیعی، استخراج اطلاعات، استخراج آزاد اطلاعات، استخراج رابطه، درخت تجزیه‌ی وابستگی

۱- مقدمه

امروزه وب جهان‌گستر بعلت توزیع‌شدگی و هزینه پایین تولید محتوا با چالش‌های جدیدی از جمله حجم زیاد اطلاعات، ناهمگنی و غیرساختاریافته بودن اطلاعات مواجه شده است. اطلاعات غیرساخت‌یافته قابل خواندن، سازماندهی و تحلیل توسط ماشین‌ها نیستند. برای اینکه بتوان از بین این حجم عظیم اطلاعات، انسان را در فهم و یافتن اطلاعات مورد نیاز یاری کرد باید بتوان متن غیرساخت‌یافته را به اطلاعات ساخت‌یافته تبدیل کرد. در واقع نیاز به سیستمی وجود دارد که بتواند داده‌ها را به شکل ساخت‌یافته درآورد. استخراج اطلاعات شامل توسعه الگوریتم‌هایی است که بصورت خودکار، متن غیرساخت‌یافته را پردازش و پایگاه داده‌ای از موجودیت‌ها، روابط و وقایع را تولید می‌کنند. استخراج روابط، اصلی‌ترین بخش استخراج اطلاعات به شمار می‌رود و در این وظیفه روابط معنایی بین موجودیت‌ها در متن کشف می‌شود. 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. روش‌های سنتی برای استخراج اطلاعات فرض می‌کنند که مجموعه‌ی ثابتی از روابط موردنظر از قبل مشخص شده‌اند. از آنجاییکه تعداد روابط موردنظر در وب بسیار بزرگ است این روش‌ها معمولا قابل گسترش به مقیاس وب نیستند، 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 [۲]. 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. اهداف کلیدی در استخراج آزاد اطلاعات عبارتنداز: (۱) مستقل از دامنه‌بودن (۲) استخراج بدون ناظر (۳) مقیاس‌پذیر بودن به حجم زیادی از متون [۱-۳].

بمنظور استخراج روابط دو نوع تحلیل زبان‌شناسی می‌تواند روی متن انجام گیرد: عمیق و سطحی. ابزار تحلیل زبان‌شناسی عمیق شامل تجزیه‌گر نحوی، تجزیه‌گر وابستگی، برچسب‌گذاری نقش معنایی، وضوح هم‌ارجاعی و.... است. ابزارهای تحلیل عمیق خودکار برای تعداد محدودی از زبان‌ها موجود است و ممکن است نتایج ناقصی را داشته باشند. تحلیل عمیقِ دستی نیز کاری دشوار، زمانبر و پرهزینه است. روش دیگر تحلیل متن، تحلیل زبان‌شناسی سطحی است که برچسب‌گذاری اجزای کلام، تجزیه‌گر سطحی، تحلیل ریخت‌شناسی و... را شامل می‌شود. ابزارهای تحلیل سطحی برای بسیاری از زبان‌ها موجود است و به اندازه‌ی کافی قابل اعتماد است [۴-۶]. ابزارهای تحلیل سطحی اغلب سریع هستند اما بدلیل محدود بودن به تحلیل سطحی باعث کاهش معیارهای کارایی نظیر دقت و بازخوانی می‌شوند [۳][۴]. ویژگی‌های مثبتِ ابزارهای سطحی و ماهیت پیکره‌های بزرگ و ناهمگنی مانند وب، استفاده از آنها را نسبت به ابزارهای عمیق باعث کاهش هزینه و در نتیجه مناسب ساخته است. یکی از اهداف این مقاله، توسعه‌ی روشی ترکیبی با در نظر گرفتن مشخصه‌های مثبت هر کدام از این رویکردها است. بکارگیری روش پیشنهادی در استخراجگرهای سطحی موجب توسعه‌ی روشی ترکیبی با در نظر گرفتن مشخصه‌های مثبت هر کدام این رویکردهای سطحی و عمیق می‌شود.

با بررسی خروجی‌های نادرست استخراجگرها، در این مقاله روشی پیشنهاد شده است که زیرمجموعه‌ای از روابط پیچیده‌ای[۳] را در نظر می‌گیرد که اغلب استخراجگرها، بویژه استخراجگرهای سطحی در کشف آنها ناتوان هستند. در این روش با کمک درخت تجزیه وابستگی و اعمال یکسری قوانین بر روی این درخت، روابط صحیح استخراج می‌شوند. این مقاله نوآوری‌های زیر را دارد :

- این مقاله به ارائه‌ی روشی جدید جهت بهبود خروجیِ سامانه‌های استخراج آزاد اطلاعات با استفاده از یک روش مبتنی بر قاعده و با کمک درخت تجزیه‌ی وابستگی می‌پردازد.
- انتخاب هوشمندانه‌ی زیرمجموعه‌ای از ورودی‌ها با کمک نوع خاصی از روابط پیچیده صورت می‌گیرد که حداقل نیاز به ابزار عمیق را دارد، در نتیجه بکارگیری روش پیشنهادی در ساختار استخراجگرهای سطحی موجب می‌شود تا روشی ترکیبی بوجود آید که از مزایای هر دو روش بهره می‌برد و در مقایسه با استخراجگرهای صرفا عمیق، سرعت بیشتر و در نتیجه مقیاس‌پذیری بیشتری دارد.
- نتایج آزمایش‌های صورت گرفته بر روی خروجی سه سامانه‌ی استخراج آزاد اطلاعات نشان می‌دهد که روش پیشنهادی می‌تواند خروجی‌های نادرست را از خروجی سامانه‌ها کاهش داده و با تبدیل آنها به استخراج‌های درست باعث بهبود معیارهای کارایی 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 توضیح مختصری از کارهای پیشین و مرتبط در زمینه‌ی استخراج آزاد اطلاعات و سامانه‌های به‌کار رفته در آزمایش‌ها ارائه خواهد شد. سامانه‌های استخراج آزاد اطلاعات را بر اساس ابزارهای بکار رفته در استخراج رابطه، می‌توان به دو دسته‌ی مبتنی بر تحلیل عمیق و سطحی تقسیم کرد. برخی استخراجگرها از ویژگی‌های سطحی مانند برچسب‌گذاری اجزا کلام [۸-۱۲] استفاده می‌کنند که استخراج سریع، در پیکره‌های مقیاس بزرگ را ممکن می‌سازد و برخی دیگر ابزارهای عمیق مانند تجزیه‌گر وابستگی [۲, ۸, ۱۳-۱۶] را بکار می‌گیرند که به هدف بهبود در معیارهای کارایی استفاده می‌شود. در ادامه تعدادی از این استخراجگرها بررسی شده است.

TextRunner [۹]: از اولین سامانه‌های استخراج آزاد اطلاعات بوده است که می‌تواند تعداد نامحدود روابط را با یک گذر در مقیاس وب استخراج

---

۱ Ontology

۲ self-supervised

۳ تعریف روابط ساده و پیچیده در [۷] بررسی شده است.

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 [۱۱]: 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 قواعد 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 [۸]: از روش خاصی برای آموزش استخراجگر که اصطلاحاً نظارت دور گفته می‌شود، استفاده می‌کند. در این سیستم از اطلاعات موجود در جعبه‌های اطلاع[۱] 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 داده‌ی آموزشی مورد استفاده قرار می‌گیرد. در واقع WOE_{POS} مثال‌های آموزشی خاصِ رابطه را با تطبیق مقادیر صفات جعبه‌های اطلاع با جملات مربوطه تولید می‌کند و این نمونه‌ها را به دادگان آموزشی مستقل از رابطه تبدیل می‌کند تا استخراجگر غیرلغوی (مستقل از لغت) یادگیری شود. سیستم WOE در دو نسخه متفاوت WOE_{POS}و WOE_{Parse} با دو سطح ویژگی ارائه شده است و کارایی بهتر از TextRunner دارد. WOE_{POS} فقط محدود به ویژگی‌های سطحی مانند برچسب‌گذاری اجزای کلام بوده و همانند TextRunner سریع است. WOE_{Parse} از ویژگی-های عمقی مانند تجزیه وابستگی استفاده می‌کند که باعث افزایش دقت و بازخوانی می‌شود. WOE_{Parse} بهترین کارایی را دارد و نشان می‌دهد که استفاده از ویژگی‌های عمیق مانند تجزیه وابستگی می‌تواند کیفیت استخراج را ارتقا دهد.

KrakeN [۱۶]: باتوجه به این که روابط باینری ممکن است شامل همه اطلاعات مورد نیاز از متن نباشد، سیستم KrakeN بر این مسئله تمرکز می‌کند، تا بتواند روابط با یک، دو و تا N آرگومان را استخراج کند. شیوه‌ی کار آن به این صورت است که ابتدا تجزیه وابستگی روی جملات انجام می‌شود. سپس عبارتی که تشخیص داده می‌شود دارای یک حقیقت است پیدا می‌شود. این عبارت زنجیره‌ای از فعل، پیرایندها و یا متمم‌هاست. در مرحله‌ی دوم رأس آرگومان‌ها توسط ارتباط‌های رو به جلو و رو به عقب در تجزیه‌گر وابستگی مشخص می‌شوند. در مرحله سوم توسط این پیوند آرگومان‌ها به صورت کامل به دست می‌آیند. این سیستم از قواعد مکاشفه‌ای استفاده می‌کند تا خطر جملاتی را که به اشتباه تجزیه شده‌اند کاهش دهد. این مسئله سبب می‌شود که بازخوانی این روش پایین باشد، همچنین استفاده از تجزیه باعث شده است که سرعت آن نسبت به سیستم‌هایی که با ویژگی‌های سطح پایین‌تر به استخراج اطلاعات دودویی می‌پردازند، پایین‌تر باشد.

ZORE [۲]: سامانه‌ی استخراج آزاد رابطه‌ی چینی است که از روش مبتنی بر نحو برای استخراج رابطه و الگوهای معنایی از متون چینی استفاده می‌کند. در این سامانه روابط کاندیدا از درخت‌های تجزیه‌ی وابستگی شناسایی می‌شوند و سپس بطور مکرر روابط به کمک الگوهای معنایی با استفاده از یک الگوریتم انتشار استخراج می‌شوند.

OLLIE [۱۴]: به هدف بهبود سامانه‌های استخراج آزاد اطلاعات توسط روشی ترکیبی و برمبنای فرایند خودراه‌انداز پیشنهاد شده است. این روش قالب‌های الگوها را بطور خودکار از یک مجموعه داده‌ی آموزشی یاد می‌گیرد که از روابط استخراج شده توسط ReVerb و با کمک فرایند خودراه‌انداز تهیه شده است. قالب الگوها از مسیر وابستگی که جفت موجودیت‌ها را بهم وصل می‌کند و روابط متناظر با آن بدست می‌آید. الگوها سپس روی پیکره اعمال می‌شوند و استخراج‌های جدید بوجود می‌آیند. OLLIE روابط n-تایی را با ترکیب روابط دودویی تولید می‌کند. OLLIE به ترتیب ۱,۹ و ۲,۷ برابر ناحیه-ی زیر نمودار بیشتری نسبت به ReVerb و WOE دارد. OLLIE از یک مولفه‌ی تحلیل محتوا استفاده می‌کند که با کمک یک فیلد اضافه سعی در تبدیل استخراج‌های نادرست به استخراج‌های درست دارد. این روش از نظر نحوه‌ی استفاده از درخت تجزیه‌ی وابستگی به روش پیشنهادی ما شباهت دارد اما علاوه بر اختلاف در نوع درخت تجزیه‌ی انتخابی، استفاده از فیلد اضافه و قواعد، در مراحل استفاده از آن نیز تفاوت عمده وجود دارد. OLLIE تجزیه‌ی وابستگی را بر تمامی ورودی‌ها اعمال می‌کند در حالیکه روش پیشنهادی در این مقاله، تجزیه‌ی وابستگی را روی زیرمجموعه‌ی کارآمدی از ورودی‌ها اعمال می‌کند که این امر موجب استفاده‌ی کمتر از ابزارهای عمیق و در نتیجه کاهش زمان و افزایش معیار مقیاس‌پذیری می‌شود که در پیکره-های بزرگی مانند وب امری بسیار مهم و حیاتی بشمار می‌رود. آزمایش‌های انجام شده در سامانه‌ی OLLIE در این مقاله بدون در نظر گرفتن مولفه‌ی تحلیل محتوای آن انجام شده است.

۳- روش پیشنهادی

با بررسی خروجی‌های نادرست استخراجگرها، روشی پیشنهاد شده است که زیرمجموعه‌ای از روابط پیچیده‌ای را در نظر می‌گیرد که اغلب استخراجگرها، بویژه استخراجگرهای سطحی از کشف آنها ناتوان هستند. یکی از روابط پیچیده مربوط به حالتی است که در آن رابطه شامل عبارت داخلی است. این نوع رابطه حالتی را نشان می‌دهد که در آن فاعل اصلی به عبارت داخلی از طریق یک فعل ارجاع دارد. این عبارت داخلی اغلب بعنوان یک مفعول برای آن فعل در نظر گرفته می‌شود [۷]. در اغلب استخراجگرها بویژه سطحی، روابط استخراجی بجای اینکه رابطه‌ی بین فاعل و مفعول را در نظر بگیرد به استخراج روابط فقط از قسمت مفعول محدود شده است. برخی جملات شامل افعال شناخت[۲] است و ممکن است منجر به چنین استخراج ناقص و نادرستی شود. برای مثال جمله‌ی زیر را در نظر می‌گیریم:

Microsoft says that Macs are too cool for its customers.

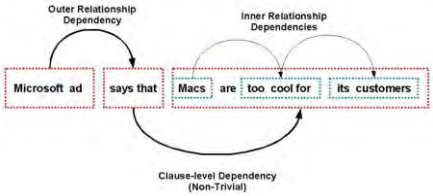
خروجی سامانه‌ی ReVerb برای چنین جمله‌ای بصورت (Macs, are too cool for, its customers) است. در شکل(۱) رابطه‌ی درونی این جمله

---

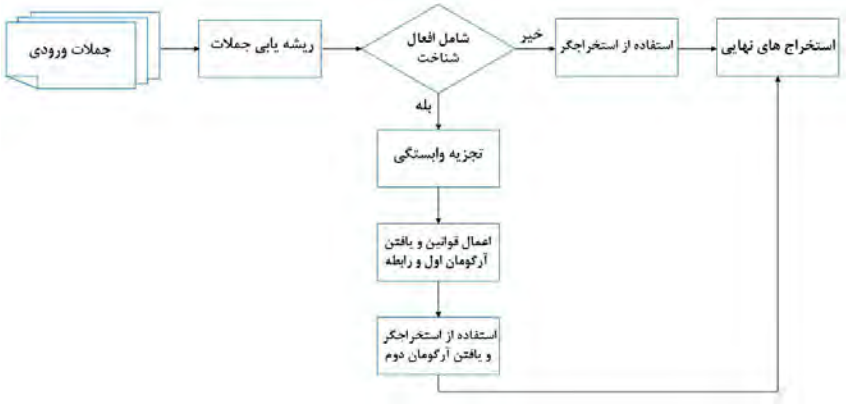
[۱] infobox

[۲] cognition

نشان داده شده است. در این استخراج ذکر نشده است که Microsoft چنین نظری را دارد و در نتیجه استخراج نادرست در نظر گرفته می‌شود. برای جلوگیری از چنین استخراج‌های نادرستی روش پیشنهادی از ساختار شکل(۲) تبعیت می‌کند که در آن ابتدا کلمات موجود در جملات ورودی ریشه‌یابی می‌شوند. سپس جملات از نظر وجود یا عدم وجود افعال شناخت بررسی می‌شوند. در واقع فعل موجود در جمله (برای مثال say) با لیستی از افعال شناختی و ارتباطی [۱۴, ۱۸] تطبیق داده می‌شوند که از VerbNet استخراج شده‌اند.



شکل (۱) : نمونه‌ای از یک رابطه با عبارت درونی[۷]



شکل(۲): ساختار کلی روش پیشنهادی

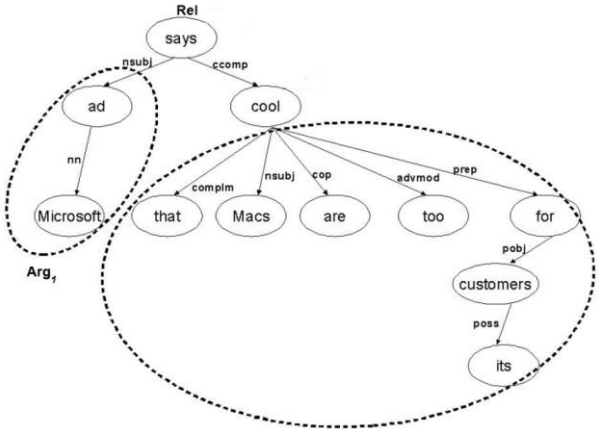
در صورت وجود این نوع افعال، روش پیشنهادی با یافتن یال ccomp (clausal complement) و اعمال قوانین، اقدام به استخراج آرگومان اول و رابطه‌ی اصلی می‌نماید.

با کمک وابستگی ccomp می‌توان رابطه‌ی بین فاعل اصلی جمله و مفعول ترکیبی را بدست آورد. در هنگام استخراج فاعل اصلی باید پیرایندهای آنرا نیز همراه با فاعل اصلی استخراج کرد که اینکار با استفاده از جستجوهای وابستگی‌های پیرایندهای مانند (nn و quant mod) ممکن می‌شود. بنابراین ساختارهای زیر استخراج خواهند شد:

Rel: {گره با دو یال بیرون‌رونده با برچسب‌های nsubj و ccomp}

Arg1: {nod1= گره‌ای که به گره Rel توسط یک یال با برچسب nsubj متصل است، گره‌ای که به گره nod1 توسط یک یال با برچسب nn یا quant mod متصل است}

خروجی ناقص ابتدایی استخراجگر برای استخراج آرگومان دوم بکار می‌رود و در نتیجه منجر به یک استخراج تودرتو[۱] می‌شود. در شکل(۳) درخت تجزیه برای جمله شکل(۱) نشان شده است و آرگومان اول و نیز رابطه‌ی اصلی استخراج شده است.



شکل(۳): کاربرد قانون استخراج برای رابطه‌ای با عبارت درونی

بنابراین با اعمال این روش خروجی درست یعنی (Microsoft ad, says, (Macs, are too cool for, its customers)) استخراج می‌شود.

علاوه بر مشکلات مربوط به موجود نبودن ابزارهای عمیق برای بیشتر زبان‌ها و دقت نامناسب این ابزارها(در مقایسه با ابزارهای سطحی) بر طبق [۱۷] بزرگ بودن مقیاس در استخراج رابطه‌ی آزاد استفاده از ابزارهای پیچیده پردازش زبان طبیعی مانند تجزیه‌گرهای نحوی و وابستگی را منع کرده است و این امر از مشکلات عمده در استخراجگرهای عمیق است. بنظر می‌رسد می‌توان با حفظ قدرت روش‌های سطحی و رفع نقایص آنها با استخراجگرها و ابزارهای عمیق و ایجاد مصامحه بین آنها به روش مناسبی برای استخراج آزاد اطلاعات دست یافت که مزایای هر دو دسته را داشته باشد. از این رو یکی از اهداف این مقاله، توسعه‌ی یک روش ترکیبی با در نظر گرفتن مشخصه‌های مثبت هر یک از استخراجگرهای سطحی و عمیق است. بکارگیری روش پیشنهادی برای بهبود کارایی و رفع مشکلاتی که اغلب در سیستم های

---

[۱] nested

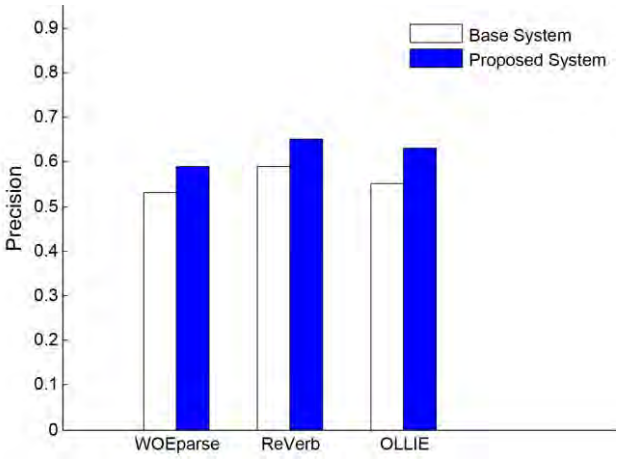
سطحی بدلیل ماهیت آنها وجود دارد موجب می‌شود تا ابزارهای عمیق روی زیرمجموعه‌ی هوشمندانه و کارامدی از ورودی‌ها اعمال شده و در نتیجه بتوان به استخراجگر ترکیبی سریع و مقیاس‌پذیری دست یافت. لازم به ذکر است که استفاده از تجزیه‌گرها نسبت به ابزارهای عمیق‌تر دیگری مانند برچسب زن نقش معنایی هزینه‌ی کمتر داشته و کارامدتر است.

٤- آزمایش‌ها و ارزیابی روش

تاثیر بکارگیری روش پیشنهادی در سامانه‌های ReVerb، WOE_{Parse} و OLLIE ارزیابی شده و رفتار آنها بررسی و مقایسه شده است. پارامترهای مختلفی وجود دارد که می‌تواند در تشخیص دقیق روابط کمک کنند. روش پیشنهادی بر طبق ساختار شکل(۲) پیاده‌سازی شده است، ابتدا جملات با کمک ریشه‌یاب استنفورد، ریشه‌یابی می‌شوند. سپس با لیست افعال شناخت موجود در [۱۴] و [۱۸] مقایسه می‌شوند و در صورت وجود این افعال، از تجزیه‌گر وابستگی استنفورد برای تجزیه‌ی این جملات استفاده شده و وابستگی ccomp در آنها جستجو می‌شود و با اعمال قوانین، خروجی‌های نهایی بدست می‌آید.

برای بررسی روش پیشنهادی از مجموعه داده‌ای استفاده شده است که در [۱۴] بکار رفته است. این مجموعه داده شامل ۳۰۰ جمله است که از سه منبع اخبار، ویکی‌پدیا و از متون زیست‌شناسی انتخاب شده است. مجموعه داده‌ی اخبار و ویکی‌پدیا در این مجموعه داده، زیرمجموعه‌ی تصادفی است که از مجموعه داده‌ی [۸] انتخاب شده است. سه سامانه‌ی ReVerb، OLLIE و WOEparse روی این مجموعه داده اجرا شده است و دو برچسب‌زن بطور دستی و مستقل هر استخراج را ارزیابی کرده‌اند و برحسب درست یا نادرست بودن خروجی‌ها برچسبی بصورت «درست» یا «نادرست» زدند. برچسب‌زن‌ها به توافق ۰,۹۶ دست یافته‌اند. آزمایش‌ها روی زیرمجموعه‌ای از داده‌ها انجام گرفته است که هر دو برچسب‌زن به توافق رسیده‌اند.

دقت بصورت نرخ تعداد استخراج‌های درست بازیابی‌شده به تعداد کل استخراج‌های بازیابی شده تعریف می‌شود. نتایج اولیه از تحلیل دقت در شکل(۴) گزارش شده است.

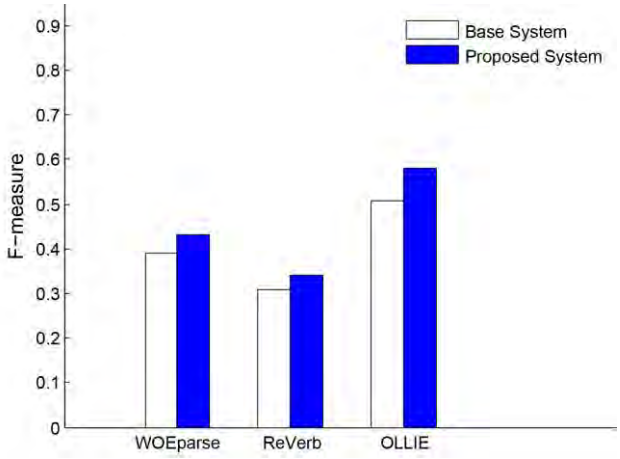


شکل(٤): مقادیر دقت حاصل در بکارگیری روش پیشنهادی در خروجی سامانه‌های استخراج آزاد اطلاعات در مقایسه با حالت پایه

نتایج آزمایش‌ها نشان می‌دهد که روش پیشنهادی می‌تواند استخراج‌های نوفه‌دار را از خروجی نهایی کاهش داده و در نتیجه باعث بهبود دقت شود. آزمایش‌های اولیه نشان می‌دهد با اعمال این روش بر روی سامانه‌های ReVerb و WOEparse حدود ۶٪ افزایش دقت حاصل می‌شود. از آنجایی که OLLIE استخراج‌های زیادی نسبت به سایر سامانه‌های مورد بررسی انجام می‌دهد، بیشترین افزایشِ دقت مربوط به سامانه‌ی OLLIE است که ۸٪ افزایش داشته است 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(۵) نتایج تحلیل مشخص شده است.

٥- نتیجه‌گیری

این مقاله به ارائه‌ی روشی جدید جهت بهبود خروجیِ سامانه‌های استخراج آزاد اطلاعات با بکارگیری قدرت ابزارهای عمیق روی زیرمجموعه‌ی کارآمدی از ورودی‌های استخراجگرها تمرکز دارد و با غنی‌کردن خروجی‌ها باعث افزایش معیارهای کارایی می‌شود. این روش زیرمجموعه‌ی خاصی از روابط پیچیده‌ای را در نظر می‌گیرد که اغلب استخراجگرها، بویژه استخراجگرهای سطحی در کشف آنها ناتوان هستند و با کمک لیست افعال شناخت، تجزیه وابستگی و با اعمال یکسری قوانین به شناسایی صحیح این نوع روابط می‌پردازد.



شکل (٥) : مقادیر امتیاز-f در بکارگیری روش پیشنهادی در خروجی سامانه‌های استخراج آزاد اطلاعات در مقایسه با حالت پایه

استخراجگرهای عمیق نسبت به استخراجگرهای سطحی برای پردازش داده‌ها از سرعت کمتری برخوردارند. این امر برای پردازش داده‌هایی در پیکره‌های بزرگ و مقیاس وب بسیار مهم و حیاتی می‌باشد و ازجمله مشکلات استخراجگرهای عمیق است. در روش پیشنهادی، با حداقل استفاده از ابزارهای عمیق روی زیرمجموعه‌ی مناسبی از ورودی‌های استخراجگرهای سطحی، پردازش داده‌های زیاد در زمان کم (در مقایسه با صرف روش‌های عمیق) ممکن است و منجر به سامانه‌ی ترکیبی می‌شود که مقیاس‌پذیری را تضمین می‌کند و کارایی بالایی دارد. نتایج آزمایش‌ها بر روی هر دو نوع استخراجگر سطحی و عمیق نشان می‌دهد که روش پیشنهادی با بهبود خروجی سامانه‌های استخراج آزاد اطلاعات باعث افزایش معیارهای کارایی می‌شود.

مراجع

[۱] رشادت وحیده، حورعلی مریم، فیلی هشام، **"ارائه روشی جهت بهبود دقت سامانه‌های استخراج آزاد اطلاعات با کمک ویژگی‌های رابطه در دامنه"**، بیست و یکمین کنفرانس ملی سالانه انجمن کامپیوتر، ۱۸۹–۱۹۴، تهران، اسفند۱۳۹۴.

[2] Qiu and Zhang. *"Zore: A syntax-based system for chinese open relation extraction"*, In Proceedings of EMNLP, 2014.

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

[4] Reshadat, Hoorali and Faili. *"A Hybrid Method for Open Information Extraction Based on Shallow and Deep Linguistic Analysis."* Interdisciplinary Information Sciences, 22:(1): 87-100, 2016.

[5] Ebadat, Claveau and Sébillot. *"Using shallow linguistic features for relation extraction in bio-medical texts."* Traitement Automatique des Langues Naturelles: 125, 2011.

[6] Bollegala, Matsuo and Ishizuka. *"Relational duality: Unsupervised extraction of semantic relations between entities on the web",* In Proceedings of the 19th international conference on World wide web, pp. 151-160, 2010.

[7] Pandit. *"Ontology-guided extraction of structured information from unstructured text: Identifying and capturing complex relationships",* 2010.

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

[9] Banko, Cafarella, Soderland, Broadhead and Etzioni. *"Open information extraction for the web"*, In IJCAI, pp. 2670-2676, 2007.

[10] Davidov, Rappoport and Koppel. "*Fully unsupervised discovery of concept-specific relationships by web mining"*, In ACL, vol. 7, pp. 232-239, 2007.

[11] Etzioni, Fader, Christensen, Soderland and Mausam. *"Open Information Extraction: The Second Generation"*, In IJCAI, pp. 3-10, 2011.

[12] Nebot and Berlanga. *"Exploiting semantic annotations for open information extraction: an experience in the biomedical domain."*, Knowledge and information Systems, 38:(2): 365-389, 2014.

[13] Gamallo, Garcia and Fernández-Lanza. "*Dependency-based open information extraction"*, In Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP, pp. 10-18, 2012.

[14] Schmitz, Bart, Soderland and Etzioni. *"Open language learning for information extraction"*, In Proceedings of EMNLP, pp. 523-534, 2012.

[15] Christensen, Soderland and Etzioni. *"An analysis of open information extraction based on semantic role labeling"*, In Proceedings of the sixth international conference on Knowledge capture, pp. 113-120, 2011.

[16] Akbik and Löser. *"Kraken: N-ary facts in open information extraction"*, In Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction, pp. 52-56, 2012.

[17] Etzioni, Banko, Soderland and Weld. *"Open information extraction from the web."* Communications of the ACM, 51:(12): 68-74, 2008.

[18] Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon*, Ph.D. Thesis, University of Pennsylvania, Pennsylvania, U.S, 2005.