بهبود هزینهٔ محاسباتی در سامانه‌های استخراج آزاد اطلاعات با استفاده از مدل لاگ‌لینیر

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

[۱] دانشکده فنی مهندسی میانه، دانشگاه تبریز، تبریز، ایران

[۲] مجتمع دانشگاهی برق و کامپیوتر، دانشگاه صنعتی مالک اشتر، تهران، ایران

[۳] دانشکده مهندسی برق و کامپیوتر، پردیس دانشکده‌های فنی دانشگاه تهران، تهران، ایران

چکیده

استخراج اطلاعات شامل توسعهٔ الگوریتم‌هایی است که به‌صورت خودکار متن غیرساخت‌یافته را پردازش و پایگاه داده‌ای از موجودیت‌ها، روابط و وقایع را تولید می‌کنند. یکی از مشکلات اساسی استخراج اطلاعات، هزینهٔ بالای محاسباتی این روش‌ها است. این موضوع در دامنه‌هایی با مقیاس بزرگ نظیر وب اهمیت زیادی دارد. در سال‌های اخیر روش‌های استخراج آزاد اطلاعات زیادی پیشنهاد شده است. این روش‌ها محدودهٔ وسیعی را از ابزارهای پردازش زبان طبیعی را اعم از سطحی (نظیر برچسب‌زن اجزای کلام) تا عمیق (نظیر برچسب‌زن نقش معنایی) در برمی‌گیرند. در این مقاله روشی بهینه برای استخراج آزاد اطلاعات نشان داده شده که بر پایهٔ ترکیب مزایای استخراج‌گرهای سطحی و عمیق و اجتناب از معایب آنها بنا شده است. استخراج‌گر که هستهٔ اصلی روش پیشنهادی است، با استفاده از پارامترهای مؤثر، زیرمجموعه‌ای را با کارایی بالا با استفاده از یک روش بهینه به کمک مدل لاگ لینیر به‌وجود می‌آورد که قابل اجرا در مقیاس وب است. این روش با بررسی جملهٔ ورودی و انتساب آن به مناسب‌ترین استخراج‌گر باعث استفاده بهینه از زمان و درنتیجه، کاهش هزینه محاسباتی شده و علاوه‌براین به‌دقت قابل قبولی نیز دست می‌یابد.

واژگان کلیدی: پردازش زبان طبیعی، استخراج اطلاعات، استخراج آزاد اطلاعات، استخراج رابطه

A New Method for Improving Computational Cost of Open Information Extraction Systems Using Log-Linear Model

Vahideh Reshadat[1]*, Maryam Hourali[2] & Heshaam Faili[3]

[1] Miyaneh Technical and Engineering Faculty, University of Tabriz, Tabriz, Iran

[2] Malek-Ashtar University of Technology, Tehran, Iran

[3]School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran

Abstract

Information extraction (IE) is a process of automatically providing a structured representation from an unstructured or semi-structured text. It is a long-standing challenge in natural language processing (NLP) which has been intensified by the increased volume of information and heterogeneity, and non-structured form of it. One of the core information extraction tasks is relation extraction which aims at extracting semantic relations among entities from natural language text. Traditional relation extraction techniques were relation-specific, producing new instances of relations determined a priori. While effective, this model is not applicable in cases where the relations are not defined a priori or when the number of relations is high. Open Relation Extraction (ORE) methods were developed to elicit instances of arbitrary relations while requiring fewer training examples. Since ORE systems are employed by the applications depended on large-scale relation

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

● تاریخ ارسال: ۹۶/۰۸/۱۸ ● تاریخ پذیرش: ۹۷/۱۰/۱۹ ● نوع مطالعه: پژوهشی

extraction, high performance and low computational cost are major requirements for ORE methods. This is particularly important in the large scales such as the Web. Many OIE systems have been proposed in recent years. These approaches range from shallow (such as part-of-speech tagging) to deep (such as semantic role labeling), therefore they differ in their performance level and computational cost.
In this paper, we use the state-of-the-art shallow NLP tools to extract instances of relations. A supervised log-linear model for OIE is presented which is based on using advantages of shallow NLP tools, as they are fast and lead to a low computational time. Extractor which is the main core of proposed approach integrates a high performance subset of the shallow NLP tools with the strength of the deep NLP tools by using a supervised log linear model and produces a high performance method that is scalable. This causes efficient use of time and therefore reduces computational cost and increases precision. Proposed approach achieves higher precision and recall than ReVerb, one of the most successful shallow OIE system.

KeyWords: Information Extraction, Open Information Extraction, Relation Extraction, Knowledge Discovery, Fact Extraction

۱- مقدمه

امروزه وب جهان‌گستر به‌علت توزیع‌شدگی و هزینه پایین تولید محتوا با چالش‌های جدیدی از جمله حجم زیاد اطلاعات، ناهمگنی و غیرساختاریافته‌بودن اطلاعات مواجه شده است. اطلاعات غیرساخت‌یافته، قابل خواندن، سازماندهی و تحلیل توسط ماشین‌ها نیستند. برای این‌که بتوان از بین این حجم انبوه اطلاعات، انسان را در فهم و یافتن اطلاعات مورد نیاز یاری کرد، باید بتوان متن غیرساخت‌یافته را به اطلاعات ساخت‌یافته تبدیل کرد. درواقع نیاز به سامانه‌ای وجود دارد که بتواند داده‌ها را به‌شکل ساخت‌یافته درآورد. استخراج اطلاعات شامل توسعه الگوریتم‌هایی است که به‌صورت خودکار، متن غیرساخت‌یافته را پردازش و پایگاه داده‌ای از موجودیت‌ها، روابط و وقایع را تولید می‌کنند. استخراج روابط[1]، اصلی‌ترین بخش استخراج اطلاعات[2] به‌شمار می‌رود و در این وظیفه روابط معنایی بین موجودیت‌ها در متن کشف می‌شود. استخراج اطلاعات نه‌تنها معنای متن را آشکار و ما را به هدف نهاییِ توانایی رایانامه‌ها به فهم متن نزدیک‌تر می‌سازد، بلکه می‌تواند در کاربردهای زیادی مانند جستجوی وب، پرسش‌وپاسخ، کاوش متون زیستی (شناسایی روابط بین پروتئین‌ها و بیماری‌ها برای کشف تأثیر جانبی بالقوه داروهای مختلف مفید است)، کسب خرد جمعی، ساخت پایگاه دانش و توسعهٔ موتورهای جستجو در یافتن نتایج مرتبط به‌کار رود. شناسایی موجودیت و استخراج رابطه، دو هسته اساسی در استخراج اطلاعات است [1-3].

نیاز به استخراج اطلاعات ساخت‌یافته از متن خام باعث به‌وجود آمدن چندین روش از جمله روش‌های مبتنی بر قالب، مبتنی بر یادگیری (باناظر، نیمه‌نظارتی و بدون ناظر) و نیز روش‌های مبتنی بر الگو و.. برای استخراج اطلاعات شده است. در ابتدا برای مدت‌های طولانی استخراج اطلاعات به‌کمک متخصصان حوزه و با روش‌های نیازمند تلاش انسانی، انجام می‌شد. این روش‌ها شامل روش‌های مبتنی بر قالب و روش‌های باناظر است که به‌ترتیب نوع حفره‌های قالب و رابطه در این روش‌ها از قبل مشخص و پرهزینه هستند و نیاز به تلاش دستی دارند. در روش‌های نیمه‌نظارتی نیز با وجود این‌که نسبت به روش‌های باناظر نیازمند دادگان برچسب‌خورده کمتری هستند؛ اما به نمونه‌های اولیه برای اجرا نیاز دارند. در سال‌های اخیر تلاش زیادی برای خودکارسازی عمل استخراج اطلاعات صورت گرفته است و در این راستا سامانه‌های استخراج رابطه بدون‌ناظر و استخراج آزاد اطلاعات[3] معرفی شدند که استخراج روابط دلخواه را از جملات در متن ممکن می‌سازند. این روش‌ها به موفقیت قابل توجهی روی پیکره‌های بزرگ و دامنه باز مانند وب دست یافته‌اند.

اهداف کلیدی استخراج اطلاعات شامل: (۱) مستقل از دامنه بودن (۲) استخراج بدون ناظر (۳) مقیاس‌پذیری با رشد تعداد متون [4]. مقیاس‌پذیری سامانه‌های استخراج آزاد اطلاعات به سطح پیچیدگی ابزارهای پردازش زبان طبیعی به‌کار رفته در آنها بستگی دارد. روش‌های استخراج آزاد اطلاعات براساس تحلیل زبانی استفاده‌شده در استخراج رابطه قابل تقسیم به دو دسته هستند. برخی از سامانه‌ها مانند تکست‌رانر [5]، ریورب [6]، WOE_{POS} [7] بر ابزارهای تحلیل نحوی سطحی[4] نظیر برچسب‌گذاری اجزای کلام و تجزیه سطحی متکی هستند. این نوع استخراج‌گرها سریع هستند؛ اما محدودبودن به تحلیل نحوی سطحی منجر به کاهش چشم‌گیر معیارهای کارایی از جمله دقت می‌شود. سایر روش‌های استخراج آزاد اطلاعات نظیر Wanderlust[8]، KrakeN[9]، DepOE[10]، OLLIE[11]، WOEparse[7]، KrakeN[9]، DepOE[31]، OLLIE[63] و WOEparse[69]

[1] Relation Extraction
[2] Information Extraction
[3] Open Information Extraction
[4] Shallow

بر تحلیل‌های معنایی و نحوی عمیق[1] مانند برچسب‌گذاری نقش معنایی و تجزیه تمرکز دارند. این استخراج‌گرها به‌طورمعمول پرهزینه‌تر از استخراج‌گرهای قبلی هستند و از طرفی کارایی بالاتری نیز دارند. استخراج‌گرهای نخست سریع هستند و از این رو مقیاس‌پذیری در وب تضمین می‌شود و نیز به‌دلیل استفاده از ویژگی‌های سطحی نیاز به تلاش کمتری دارند. این استخراج‌گرها از مدیریت ساختارهای پیچیده‌ای مانند شناسایی روابط راه دور ناتوان هستند و به‌دلیل استفاده از ویژگی‌های سطحی، کارایی پایینی دارند؛ درحالی‌که استخراج‌گرهای دستهٔ دوم به‌دلیل استفاده از ابزارهای تحلیل معنایی یا نحوی عمیق نظیر تجزیه‌گر وابستگی[2] از نظر کارایی بهتر، ولی زمان‌گیر بوده و به‌دلیل استفاده از ویژگی‌های عمیق در فرایند استخراج، پرهزینه هستند و مقیاس‌پذیری به‌نسبه پایین‌تری دارند.

با داشتن مزایا و معایب هر کدام از استخراج‌گرهای سطحی و عمیق، در این مقاله روشی برای تخمین خودکار دشواری جملهٔ ورودی به سامانه‌های استخراج آزاد اطلاعات پیشنهاد شده است. برای این‌کار از یک دسته‌بند رگرسیون لجستیک استفاده شده است که جملاتی را که برای استخراج‌گرهای سطحی دشوار هستند به استخراج‌گرهای عمیق انتقال می‌دهد. بنابراین روش پیشنهادی جملات را با هدف کاهش هزینه محاسباتی طبقه‌بندی می‌کند. روش پیشنهادی، ترکیبی از دو نوع از سامانه‌های استخراج آزاد اطلاعات است که از ریورب[3] [12] و اِگزمپلر[4] [13] به‌عنوان استخراج‌گرهای سطحی و عمیق به‌ترتیب استفاده می‌کند.

در این مقاله مسامحه بین کارایی(معیار-f) و هزینهٔ محاسباتی بررسی شده است. نتایج نشان می‌دهد که به‌کاربردن استخراج‌گر عمیق به روی زیرمجموعهٔ هوشمندی از جملات ورودی می‌تواند بهبود قابل توجهی در معیار-f حاصل کند. این مقاله نوآوری‌ها و دستاوردهای زیر را دارد:

- روشی نوین برای پیش‌گویی دشواری جملات در استخراج آزاد رابطه به کمک دسته‌بند رگرسیون لجستیک نشان داده شده است که جملات دشوار را برای استخراج رابطه توسط استخراج‌گرهای سطحی به استخراج‌گر عمیق هدایت می‌کند؛ بنابراین دسته‌بندِ دشواری، جملات را براساس احتمال بهبود کارایی اولویت‌بندی می‌کند.
- هدف اصلی روش پیشنهادی به‌دست‌آوردن کارایی بالا در زمان محاسباتی پایین است. با به‌کاربردن استخراج‌گرهای عمیق برای جملاتی که استخراج‌گرهای سطحی قادر به استخراج صحیح از آنها نیستند، این روش قادر به تخصیص بهتر منابع محاسباتی (دستِ‌کم استفاده از ابزارهای عمیق پردازش زبان طبیعی) و اجتناب از اتلاف این منابع در جملاتی است که احتمال بهبود کارایی در آنها کم است. این روش از زمان موجود، استفادهٔ مؤثر می‌کند؛ علاوه‌براین، دسته‌بند دودویی به‌گونه‌ای آموزش داده می‌شود که نمونه‌های صحیح بیشتر و نمونه‌های غیرصحیح کمتری را استخراج کند؛ بنابراین کارایی (معیار-f) افزایش می‌یابد.
- از مجموعه‌ای از ویژگی‌های سبک‌وزن استفاده می‌شود که دانش زبانی سطحی را به دسته‌بند اعمال می‌کنند. همهٔ ویژگی‌ها به‌صورت کارا قابل محاسبه بوده و مستقل از رابطه هستند. این ویژگی‌ها مقیاس‌پذیر، مستقل از دامنه و در سطح جمله هستند و بنابراین می‌تواند به‌راحتی در زمان استخراج بدون استفاده از ابزارهای عمیق محاسبه شوند.
- آزمایش‌ها نشان می‌دهد که ترکیب استخراج‌گرهای آزاد اطلاعات سطحی و عمیق و پیداکردن بهترین مسامحه بین سود و زیان آنها می‌تواند توازن بالایی از دقت و بازخوانی را در مقایسه با استخراج‌گر سطحی تشکیل‌دهندهٔ آن به‌وجود آورد. این مقدار در مقایسه با استخراج‌گر عمیق تشکیل‌دهندهٔ آن یکسان بوده، اما در زمانی بسیار کمتر حاصل می‌شود؛ بنابراین در شرایطی مفید است که مجموعهٔ داده بزرگ بوده و زمان پردازش محدود باشد؛ بنابراین این روش گام امیدبخشی را برای استخراج آزاد رابطه مقیاس‌پذیر فراهم می‌کند.

ادامهٔ مقاله به‌صورت زیر سازماندهی شده است. در بخش دو کارهای پیشین در زمینهٔ سامانه‌های استخراج‌گرهای آزاد اطلاعات معرفی و مفهوم جملات دشوار در استخراج اطلاعات در بخش سه بررسی و روش پیشنهادی در بخش چهار شرح داده شده است. نتایج آزمایش‌ها در بخش پنج نشان داده می‌شود و با نتیجه‌گیری در بخش شش پایان می‌یابد.

۲- کارهای مرتبط

در این بخش تعدادی از کارهای مرتبط در استخراج آزاد اطلاعات به‌ویژه کارهای مرتبط با استخراج‌گر آزاد رابطه بررسی شده است. سطوح مختلفی از ابزارهای پردازش زبان طبیعی از سطحی (مانند برچسب‌گذاری اجزای کلام) تا عمیق

[1] Deep
[2] Dependency Parser
[3] ReVerb
[4] Exemplar

(برچسب‌زن نقش معنایی) در استخراج‌گرهای آزاد اطلاعات به‌کار رفته شده است [1,4,10]. این سامانه‌ها براساس تحلیل زبانی به‌کاررفته در عمل استخراج می‌توانند به چند دستهٔ اصلی تقسیم شوند که در ادامه بررسی شده است.

۱-۲- روش‌های استخراج آزاد اطلاعات سطحی

در این بخش روش‌های استخراج آزاد اطلاعات سطحی بررسی می‌شوند که بر ابزارهای سطحی زبان طبیعی تکیه می‌کنند. سامانه تکست رانر[1] [14] از نخستین سامانه‌های استخراج آزاد اطلاعات بوده است که می‌تواند تعداد نامحدود روابط را با یک گذر در مقیاس وب استخراج کند. این سامانه مستقل از دامنه است و یک رابطه و آرگومان‌های آن را بدون استفاده از الگوهای واژگانی و با روش خودناظر استخراج می‌کند. از آنجایی‌که استفاده از تجزیه‌گر[2] برای استخراج روابط در مقیاس وب عملی نیست، می‌توان از آن برای آموزش استخراج‌گر استفاده کرد. تجزیه‌گر به استخراج‌گر کمک می‌کند تا مجموعه‌ای از روابط مورد اطمینان را شناسایی و برچسب بزند. برای ایجاد مجموعه آموزشی برای هر جملهٔ تجزیه‌شده، سامانه عبارت‌های اسمی و ساختار تجزیه‌ای را که دو موجودیت را به‌هم وصل می‌کند پیدا می‌کند و با اعمال محدودیت‌هایی در صورت برقراربودن محدودیت‌ها (احتمال درست‌بودن آنها زیاد است) برچسب مثبت می‌خورند و درصورتی‌که هر کدام از این شرایط برقرار نباشد، برچسب منفی در نظر گرفته می‌شود. سامانه تکست‌رانر از داده‌هایی که خودش برچسب‌زده است، استفاده می‌کند، تا عبارت‌های رابطه‌ای[3] را بیابد و یک مدل از نوع دسته‌بند که مشخص‌کنندهٔ وجود یا عدم وجود رابطه است، تولید می‌کند. برای بهره‌گیری از مدل تکست‌رانر و اعمال آن روی متن ورودی به منظور استخراج اطلاعات، ابتدا موجودیت‌های نامدار در متن شناسایی و سپس رابطهٔ بین هر دو موجودیت توسط دسته‌بندِ آموزش دیده تشخیص داده می‌شود.

ریورب [12] یک سامانهٔ موفق و قدرتمند سطحی برای استخراج اطلاعات است. این سامانه از دنباله‌ای از برچسب‌گذاری‌های اجزای کلام به‌عنوان یک محدودیت نحوی استفاده می‌کند تا عبارات رابطه‌ای را استخراج و استخراج‌های غیرمنسجم و استخراج‌هایی را که اطلاعی در برندارند حذف کند. اگر برای یک فعل چند انطباق با قواعد نحوی در یک جمله وجود داشته باشد، طولانی‌ترین آنها انتخاب می‌شود و در مواردی که انطباق‌های متوالی وجود داشته باشد، این انطباق‌ها با هم ترکیب می‌شوند؛ که یک رابطه بزرگ‌تر تشکیل شود. ممکن است، عبارات رابطه‌ای وجود داشته باشند که محدودیت نحوی را ارضا کنند؛ ولی بسیار خاص باشند. برای غلبه بر این مشکل یک محدودیت لغوی استفاده می‌کند که هدفش کاهش تعداد استخراج‌های بسیار خاص است. این محدودیت بر این اساس است که یک عبارت رابطه‌ای معتبر باید تعداد زیادی آرگومان مجزا در پیکرهٔ بزرگ داشته باشد. ریورب از یک دسته‌بند احتمالاتی استفاده می‌کند که برای هر خروجی یک ضریب اطمینان تخصیص می‌دهد. آزمایش‌ها نشان می‌دهد که ریورب بهتر از تکست‌رانر عمل می‌کند و کارایی آن دو برابر تکست‌رانر است.

در [15] سامانه SONEX پیشنهاد شده که توسعه‌یافتهٔ ریورب است. این روش مجموعه جفت‌موجودیت‌ها را می‌گیرد و رابطهٔ مربوطه را تولید می‌کند. خروجی این سامانه جفت‌های متعلق به یک رابطه است که داخل یک خوشه هستند و یک برچسب دارند. درواقع این الگوریتم با خوشه‌بندی‌کردن جفت موجودیت‌ها کار می‌کند. برای جفت‌موجودیت در جمله، حداکثر N کلمه بینشان انتخاب می‌شود و به‌صورت برداری از ویژگی‌ها نشان داده می‌شوند. این ویژگی‌ها شامل یک‌تایی[4]، دوتایی[5] و برچسب اجزای کلامِ کلمات بین دو جفت مجودیت است. از معیار شباهت کسینوسی برای محاسبه میزان شباهت بین بردارها در عمل خوشه‌بندی استفاده می‌شود. در این مقاله روش وزن‌دهی جدیدی پیشنهاد شده است تا قدرت تمیز یک واژه داخلِ دامنه رابطه را محاسبه کند. از این روش وزن‌دهی در ساخت ماتریس کلمات استفاده می‌شود. برای خوشه‌بندی بردارهای حاصل، از الگوریتم خوشه‌بندی سلسله‌مراتبی استفاده شده است؛ زیرا نیازی به تعیین تعداد خوشه‌ها ازقبل ندارد. برای اینکه بتوان از این الگوریتم در مقیاس بزرگ استفاده کرد، در اینجا از الگوریتم باک‌شات[6] استفاده شده است که چون نمونه‌ای از بردارها را خوشه‌بندی می‌کند، پیچیدگی زمان و هزینه را کاهش می‌دهد. نتایج ارزیابی این سامانه بهتر از سامانه استخراج آزاد اطلاعات ریورب بوده است.

سامانه WOE[7] [7] نیز از روش خاصی برای آموزش استخراج‌گر استفاده می‌کند که اصطلاحاً نظارت دور گفته می‌شود. در این سامانه از اطلاعات موجود در جعبه‌های اطلاع[8]

[1] TextRunner
[2] Parser
[3] Relational phrase

[4] unigram
[5] bigram
[6] Buckshot
[7] Wikipedia-based Open Extractor (WOE)
[8] infobox

ویکی‌پدیا استفاده می‌شود. هر اطلاع یک رابطهٔ دوتایی است که یکی از آرگومان‌های آن موضوع صفحهٔ ویکی‌پدیا و دیگری مقادیر صفات آن است. با انطباق اطلاعات با جملات متن، جملات و رابطهٔ استخراج‌شده از آن‌ها به‌دست می‌آید و به‌عنوان داده آموزشی مورد استفاده قرار می‌گیرد. درواقع WOE مثال‌های آموزشی خاص-رابطه را با تطبیق مقادیر صفات جعبه‌های اطلاع با جملات مربوطه تولید می‌کند؛ اما WOE این نمونه‌ها را به دادگان آموزشی مستقل از رابطه تبدیل می‌کند تا استخراج‌گر غیرلغوی (مستقل از لغت) یادگیری شود. سامانه WOE در دو نسخه متفاوت WOE_{POS}و WOE_{Parse} با دو سطح ویژگی ارائه شده است و کارایی بهتر از تکسترانر دارد. WOE_{POS} فقط محدود به ویژگی‌های سطحی مانند برچسب‌گذاری اجزای کلام بوده و همانند تکسترانر سریع است و WOE_{Parse} از ویژگی‌های 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 [1].

سامانهٔ R2A2 [6] با توسعهٔ سامانهٔ ریورب در بخش استخراج آرگومان‌های رابطه تولید شده و بسیاری از خطاهای آن را که در اثر آرگومان‌های اشتباه بوده، بهبود داده است. در سامانه ریورب آرگومان‌های روابط با استفاده از چند قانون مکاشفه‌ای استخراج می‌شود؛ اما در R2A2 آرگومان‌ها با استفاده از دسته‌بند CRF آموزش داده می‌شوند. یادگیر آرگومان‌ها به دو بخشِ یادگیری محدوده راست و چپ هر آرگومان تقسیم می‌شود. این یادگیر آرگومان از سه دسته‌بند برای این منظور استفاده می‌کند که دو دسته‌بند برای شناسایی محدوده راست و چپ آرگومان نخست و یکی نیز برای برای شناسایی محدوده راست آرگومان دوم به‌کار می‌رود. از آنجایی‌که آرگومان دوم به‌طورتقریبی همیشه به‌دنبال عبارت رابطه‌ای می‌آید نیاز به دسته‌بند جداکنندهٔ محدوده چپ آرگومان دوم وجود ندارد. ارزیابی‌ها نشان می‌دهد R2A2 شناسایی آرگومان در مقایسه با قواعد مکاشفه‌ای مربوط به ریورب بهبود پیدا کرده و باعث کاهش خطا در تشخیص هر دو آرگومان رابطه شده است.

۲-۲- روش‌های استخراج آزاد اطلاعات عمیق

روش‌های استخراج آزاد اطلاعات عمیق که از ابزارهای عمیق زبان طبیعی استفاده می‌کنند، در این بخش بررسی می‌شوند.

ZORE [16] یک سامانهٔ استخراج رابطهٔ مبتنی بر نحو برای استخراج روابط و الگوهای معنایی از متون چینی است. این سامانه نامزدهای روابط را به‌طور خودکار از درخت‌های تجزیه وابستگی شناسایی و سپس توسط یک الگوریتم انتشار، روابط با الگوهای معنایی‌شان را به‌طور تکراری استخراج می‌کند. روش [17] نیز روی استخراج آزاد روابط چینی تمرکز دارد. این سامانه به‌صورت خط لوله‌ای از وظایف شامل قطعه‌بندی کلمات، برچسب‌زنی اجزای کلام و تجزیه است.

سامانه Dep-OE [10] نیز از ویژگی‌های تجزیه در استخراج روابط استفاده می‌کند. این سامانه از هر جمله چند حقیقت در قالب روابط دوتایی استخراج می‌کند و به استخراج اطلاعاتی که با اجزایی غیر از فعل بیان می‌شود، نیز توجه دارد. همچنین به استخراج خردجمعی از حقایق نیز توجه داشته است و گزاره‌های پایه‌ای را از متن استخراج می‌کند. در این روش یک استخراج‌گر آزاد اطلاعات چندزبانه براساس تجزیه‌گر وابستگی مبتنی بر قاعده پیشنهاد شده که روشی سریع و مقاوم است. این روش شامل سه مرحله است: هر جمله متن ورودی با استفاده از تجزیه‌گر DepPattern تحلیل می‌شود. این تجزیه‌گر شامل گرامر برای پنج زبان و نیز کامپایلر برای ساخت تجزیه‌گرها در پرل[1] است؛ سپس برای هر جملهٔ تجزیه‌شده عبارات فعلی آن تشخیص داده شده و سپس برای هر عبارت اجزای فعل (شامل فاعل، مفعول مستقیم، صفت و مکمل‌های اضافی) شناسایی می‌شوند و درنهایت مجموعه‌ای از قوانین به اجزای عبارت که در مرحلهٔ قبلی شناخته شده‌اند، به‌منظور استخراج سه‌تایی‌های موردنظر اعمال می‌شود. قوانین استفاده‌شده در این روش برای استخراج سه‌تایی‌ها، مبتنی بر فعل هستند و نیز فقط یک سه‌تایی از یک عبارت استخراج می‌کنند. این سامانه در مقایسه با سامانهٔ استخراج آزاد اطلاعات ریورب دقت و نیز معیار-F بالاتری دارد. نسخهٔ جدیدی از DepOE به‌نام ArgOE در [18] پیشنهاد شده است. ArgOE یک روش استخراج آزاد اطلاعات مبتنی بر قاعده است که تجزیه‌های وابستگی به شکل CoNLL-X را به‌عنوان ورودی می‌گیرد و ساختار آرگومان‌های داخل تجزیه‌های وابستگی را شناسایی کرده و مجموعه‌ای از گزاره‌ها را از ساختار آرگومان استخراج می‌کند. این روش نیازی به دادگان آموزشی ندارد و دقت و بازخوانی بالایی نسبت به روش‌های پیشینی دارد که به دادگان آموزشی وابسته هستند.

LSOE [19] یک استخراج‌گر بر مبنای الگوهای لغوی-نحوی است که دو نوع الگو استخراج می‌کند: ۱-الگوهای کلی

[1] Perl

۲- قواعدی از روش پیشنهادی در [20]. ایدهٔ اصلی این است که یک راه‌حل ساده‌ای برای اجرای استخراج مبتنی بر قاعده‌ای از سه‌تایی‌ها از متنی صورت گیرد که برچسب‌گذاری اجزای کلام در آن انجام گرفته است. کارایی LSOE با دو سامانهٔ استخراج آزاد اطلاعات ریورب و DepOE مقایسه شده است. نتایج نشان می‌دهد که LSOE روابطی را به‌درستی استخراج می‌کند که توسط دیگر استخراج‌گرها یادگیری نشده است و درنتیجه به دقت قابل ملاحظه‌ای دست یافته می‌یابد.

اغلب روش‌های استخراج آزاد اطلاعات محدود به روابط دودویی هستند؛ برخی از روش‌ها به استخراج روابط n-تایی نیز می‌پردازند. باتوجه به این که روابط دودویی ممکن است، شامل همه اطلاعات مورد نیاز از متن نباشد، سامانه KrakeN [9] بر این مسأله تمرکز می‌کند، تا بتواند روابط با یک، دو و تا N آرگومان را استخراج کند. شیوهٔ کار آن به این صورت است که ابتدا تجزیه وابستگی روی جملات انجام و سپس عبارتی پیدا می‌شود که تشخیص داده شود دارای یک حقیقت است. این عبارت زنجیره‌ای از فعل، پیرایندهها و یا متمم‌هاست. در مرحلهٔ 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 به‌دست می‌آیند.

این سامانه از قواعد مکاشفه‌ای استفاده می‌کند تا خطر جملاتی را کاهش دهد که به‌اشتباه تجزیه شده‌اند. بنابراین از جملاتی عبور می‌کند که احتمال اشتباه در تجزیه آنها وجود دارد. این مسئله سبب می‌شود که بازخوانی این روش پایین باشد؛ همچنین استفاده از تجزیه باعث شده است که سرعت آن نسبت به سامانه‌هایی که با ویژگی‌های سطح پایین‌تر به استخراج اطلاعات دودویی می‌پردازند، پایین‌تر باشد.

OLLIE [21] یک استخراج‌گر اطلاعات است که از روش‌های یادگیری ماشین استفاده می‌کند و قالب‌های الگوی آزاد را از روی مجموعه آموزش یاد می‌گیرد. این روش یک رویکرد ترکیبی براساس روش خودراه‌انداز است که قالب‌های الگو را به‌طور خودکار از مجموعه دادهٔ آموزشی یاد می‌گیرد که از روابط استخراج‌شده توسط ریورب استفاده می‌کند. مسیرهای وابستگی که جفت‌موجودیت‌ها و روابط مرتبط با آنها را به‌هم وصل می‌کند، قالب‌های الگوها را برای OLLIE تولید می‌کنند. الگوها سپس روی پیکره اعمال می‌شود و حقایق جدید به‌دست می‌آید. OLLIE استخراج‌گرهای n-تایی را توسط ادغام روابط دودویی تولید می‌کند. مساحت زیر نمودار برای نمودارهای دقت در این سامانه حدود ۱,۹ تا ۲,۷ برابر در مقایسه با ریورب و woe است.

Patty [22] الگوهای متنی را از جمله‌ها براساس مسیرهای موجود در درخت تجزیهٔ وابستگی بین دو موجودیت اسمی استخراج می‌کند. برای تمام جفت‌های موجودیت اسمی، patty کوتاه‌ترین مسیر را در درخت وابستگی پیدا می‌کند که دو موجودیت اسمی را به‌هم وصل می‌کند. این روش جستجو را فقط به مسیرهایی محدود می‌کند که با یکی از یال‌های وابستگی خاص شروع می‌شود.

ClauseIE [4] یک روش مبتنی بر عبارت است که از روش‌های دیگر از این جهت متفاوت است که شناسایی قسمت‌های مفید از اطلاعات بیان‌شده در جمله را از نمایش آنها در استخراج‌ها جدا می‌کند. این روش از دانش زبانی در گرامر زبان انگلیسی استفاده می‌کند تا عباراتی را از جملات ورودی شناسایی کند. برای این‌کار نوع هر عبارت را مطابق تابع گرامری اجزای آن شناسایی می‌کند. این روش بر اساس تجزیهٔ وابستگی و مجموعهٔ کوچکی از لغات مستقل از دامنه است و جملات را بدون پس‌پردازش بررسی می‌کند و نیاز به دادهٔ آموزشی ندارد. این روش دقت و بازخوانی بالایی دارد و می‌تواند برای استخراج روابط n-تایی نیز به‌کار رود.

اِگزمپلر [13] مسألهٔ استخراج روابط n-تایی را به‌کمک قواعد دست‌نویس روی درخت‌های وابستگی مورد خطاب قرار می‌دهد. این قواعد روی هر آرگومان نامزد با جستجوی مسیر بین یک موجودیت و کلمهٔ رابطه‌ای به‌تنهایی اعمال می‌شود. از آنجایی که هدف دست‌یابی به دقت بالا و هزینهٔ محاسباتی پایین است، حالات مختلفی از این روش توسط تجزیه‌گرهای وابستگی مختلفی نشان داده شده است. نتایج امیدبخش است و این روش از سامانه‌های دیگری که برای استخراج روابط n-تایی به‌کار می‌رود کارایی بهتری دارد؛ درحالی‌که زمان محاسباتی پایینی دارد.

CSD-IE [23] روشی است که از تجزیهٔ جملهٔ متنی برای استخراج آزاد اطلاعات استفاده می‌کند. یک جمله به قسمت‌هایی تجزیه می‌شود که از نظر معنایی به‌هم وابسته هستند و سپس فعل(صریح یا ضمنی) در هر قسمت شناسایی می‌شود و حقایق به‌دست می‌آیند. این روش با ریورب، Ollie و clauseIE به‌کمک سه ویژگی مهم مقایسه شده است. استخراج‌های صورت‌گرفته به‌کمک قواعد استنباطی در [24] غنی می‌شوند. نتایج ارزیابی‌ها نشان می‌دهد این روش تعداد استخراج‌های صحیح و حاوی اطلاعات مفید را تا ۱۵٪ افزایش

می‌دهد که این کار را با کاهش استخراج‌های غیرمفید انجام می‌دهد.

R-OpenIE [25] یک روش استخراج آزاد اطلاعات مبتنی بر قاعده است که از یک مبدل وضعیت-متناهی استفاده می‌کند. این روش قواعد اعلانی محدود متنی برای تولید الگوهای استخراج رابطه تعریف می‌کند و از مدل مبدل وضعیت-متناهی آبشاری برای تطبیق تاپل‌های رابطه‌ای استفاده می‌کند که شرایط را ارضا می‌کنند. این روش در طول فرایند تطبیق مبدل وضعیت-متناهی آبشاری، برای هر وضعیت تطبیق داده شده، اندیس معکوس ایجاد می‌کند. بنابراین کارایی تطبیق الگو بهبود می‌یابد.

روش TreeKernel [26] نخست بررسی می‌کند که آیا یک رابطه بین یک جفت از موجودیت‌ها در یک جمله وجود دارد و سپس وجود و یا عدم وجود کلمات رابطه‌ای صریح برای این جفت را بررسی می‌کند. تعدادی مدل ماشین بردار پشتیبانی با کرنل‌های درخت وابستگی به‌کار گرفته شده است. مجموعه‌ای از الگوهای نحوی برای تولید روابط نامزد استفاده شده است؛ سپس نمونه‌های نامزد با تلفیق همهٔ جفت موجودیت‌ها با روابط استخراج‌شده از یک جمله تولید می‌شوند. با داشتن نمونهٔ نامزد، tree kernel تعدادی مسیر را با به‌کاربردن درخت وابستگی استخراج می‌کند. یک دسته‌بند هسته درخت سپس یک نمونه رابطهٔ صحیح را شناسایی می‌کند. اگرچه این سامانه بهتر از ریورب و OLLIE عمل می‌کند، اما به زمان محاسباتی بالایی نیاز دارد.

۳-۲- روش‌های استخراج آزاد اطلاعات ترکیبی

در این بخش روش‌های استخراج آزاد اطلاعاتی بررسی خواهد شد که از ترکیب روش‌های سطحی و عمیق استفاده می‌کنند. [27] استفاده از سامانه‌های برچسب‌گذاری نقش معنایی[1] در استخراج آزاد روابط بررسی شده است. فعل و آرگومان‌هایش مطابق با رابطه و آرگومان‌های رابطه است. برچسب نقش معنایی، اطلاعاتی بیشتری از آنچه که استخراج آزاد اطلاعات لازم دارد، فراهم می‌کند. بنابراین برای مقایسهٔ بهتر با سامانه‌های استخراج آزاد اطلاعات، خروجیِ برچسب‌گذار نقش معنایی به استخراج‌ها تبدیل می‌شود. سامانهٔ برچسب‌گذار نقش معنایی اجزای جمله را با درنظرگرفتن فعل برچسب می‌زند. در این رویکرد از سامانه‌های UIUC-SRL [28] و LUND-SRL [29] به‌عنوان سامانه‌های پایه برای برچسب‌زنی نقش معنایی استفاده شده است. قدرت و ضعف استخراج‌گرها در پیکرهٔ کوچک که در آن زمان محاسباتی زیادی وجود دارد و پیکرهٔ بزرگ که سامانه‌ها نمی‌توانند فرآیند را تکمیل کنند، بررسی شده است. در پیکرهٔ کوچک، سامانه SRL-IE-LUND بالاترین دقت و SRL-IE-UIUC بالاترین بازخوانی و بالاترین F1 را دارد. با تغییر در خروجی این روش دو سامانه PRECHYBRID و PRECHYBRID به‌وجود آمده است.

درصورتی‌که اجتماع خروجی سامانه‌های مبتنی بر برچسب‌گذار نقش معنایی و زیرمجموعه‌ای از تکسترانر که بالاترین دقت را دارد، به‌دست آورده شود، بالاترین بازخوانی و شاخص-F به‌دست می‌آید. زیرمجموعه‌ای از تکسترانر که بالاترین دقت را دارد توسط معیار رتبه‌بندی محلی[2] جدید پیشنهادشده، تعیین می‌شود. اگرچه SRL-IE-LUND بالاترین دقت را دارد، شرایطی وجود دارد که تحت آن شرایط تکسترانر دقت بالاتری را به‌دست می‌آورد. کارایی این سامانه‌ها تحت دو روش رتبه‌بندی مختلف، یکی محلی‌بودن و دیگری افزونگی[3] بررسی شده است.

افزونگی تعداد دفعاتی است که رابطه از جملات مختلف استخراج می‌شود. نتایج رتبه‌بندی برحسب افزونگیِ استخراج‌های دودویی نشان می‌دهد که افزونگی دقت تکسترانر را بهبود می‌بخشد، اما بازخوانی را به‌شدت کاهش می‌دهد. برای پیکره‌های با افزونگی بالا تکسترانر الگوریتم مناسبی دارد؛ اما آزمایش‌ها به‌وضوح نشان می‌دهد استخراج‌های با فراوانی افزونگی بالا در مقیاس وب محدود است. محلی‌بودن، تعداد توکن‌های بین نخستین و آخرین آرگومان در جمله است. این معیار به هردو سامانه‌ها کمک می‌کند دقت بالاتری در بازخوانی بالاتری نسبت به بازخوانی مربوط به افزونگی به‌دست آورند.

با درنظرگرفتن محلی‌بودن و افزونگی می‌توان زیرمجموعه‌ای از تکسترانر که بالاترین دقت را دارد، به‌دست آورد که برای استخراج‌های دودویی روابطی که معیار محلی‌بودن را داشته باشند، حذف می‌شوند. به‌طور کلی استخراج‌گرهای مبتنی بر برچسب‌گذاری نقش معنایی به‌خوبی عمل می‌کنند. هرچند تکسترانر تحت رتبه‌بندی محلی‌بودن، دقت بالاتری در بازخوانی بالاتری دارد، هیچکدام از افزونگی و محلی‌بودن استخراج‌گرهای مبتنی بر برچسب‌گذاری نقش معنایی را بهبود نداده (به‌جز برای nتایی). عیب اصلی استخراج‌گرهای مبتنی بر برچسب‌گذاری نقش معنایی این است که بارها کندتر از تکسترانر است.

[1] SRL: Semantic Role Labeling

[2] locality

[3] redundancy

نتایج نشان می‌دهد، استخراج‌گرهای مبتنی بر برچسب‌گذاری نقش معنایی نسبت به متن وب که ناخالصی دارد، مقاوم است و به بازخوانی بیشتری دست می‌یابد؛ درحالی‌که تکست رانر به دقت بالاتری دربازخوانی پایین‌تری نسبت به استخراج‌گر مبتنی بر برچسب‌گذاری نقش معنایی دست یافته است. سرانجام تکست‌رانر ۲۰-۷۰۰ برابر سریع‌تر از سامانه‌های استخراج‌گر مبتنی بر برچسب‌گذاری نقش معنایی آزمایش شده است (این امر بستگی به استفاده از تجزیه وابستگی یا سازه‌ای دارد).

EFFICIENS [30] از دو پارامتر آلفا و بتا استفاده می‌کند که مقادیرشان در بازهٔ [۰، ۱] است. این پارامترها تعداد جملاتی را تعریف می‌کنند که می‌توانند توسط ابزارهای پرهزینهٔ پردازش زبان طبیعی (تجزیه وابستگی و برچسب‌زن نقش معنایی) پردازش شود. با داشتن مجموعه‌ای از جملات، آلفا تعریف می‌کند که به اندازهٔ نسبت آلفا از تعداد کل جملات باید توسط تجزیه‌گر وابستگی پردازش شود؛ درحالی‌که به اندازهٔ نسبت بتا از این تعداد نیز باید توسط برچسب‌گذار نقش معنایی پردازش شود. در ابتدا تمام جملات توسط برچسب‌گذار اجزای کلام پردازش می‌شوند. این روش برای هر ابزار پردازش زبان طبیعی یک ماژول دارد. ماژول Efficiens[pos] به برچسب‌گذاری اجزای کلام تکیه دارد؛ درحالی‌که efficiens [dep] و efficiens[srl] به‌ترتیب به تجزیه‌گر وابستگی و برچسب‌گذار نقش معنایی تکیه دارد. هر ماژول می‌تواند به‌تنهایی به‌عنوان روش ORE بکار رود.

TR-DOE و RV-DOE [1] دو سامانه ترکیبی هستند که زیرمجموعه با کارایی بالا از سامانه استخراج آزاد اطلاعات سطحی را با قدرت سامانه استخراج آزاد اطلاعات عمیق ترکیب می‌کند. بهترین مسامحه بین دقت و بازخوانی با تنظیم دو پارامتر ترکب طول جمله و معیار ضریب اطمینان شناسایی شده است. از آنجایی که تمرکز بر استفادهٔ بهینه از زمان است؛ از یک استخراج‌گر عمیق سریع و قوی استفاده شده است. آزمایش‌ها نشان می‌دهد که روش‌های ترکیبی پیشنهادی، کارایی بالایی از سامانه‌های تشکیل‌دهنده‌شان دارند. بهترین نتیجه برای TR-DOE است که معیار- F به‌طورتقریبی دو برابر تکست‌رانر دارد. یکی از معایب عمدهٔ این روش‌ها کمبود عامل‌های مؤثر است. به عبارت دیگر، عامل‌های به‌کاررفته برای ترکیب بسیار خاص هستند. برای مثال، روش مبتنی بر ضریب اطمینان برای ترکیب روش‌ها، فقط از امتیاز اطمینان استخراج‌گرهای سطحی استفاده می‌کند. بنابراین برای استخراج‌گرهای سطحی که فاقد ضریب اطمینان هستند، نمی‌تواند به‌کاررود. علاوه‌براین، روش یادشده، استخراج‌گر سطحی را برای تمام جملات ورودی اجرا می‌کند و سپس از بین آنها جملات منتخبی را برای پردازش توسط استخراج‌گر عمیق به‌کار می‌برد. روش‌های مبتنی بر طول جمله نیز علاوه‌بر محدودبودن به طول جمله، برای جملاتی از ورودی که طول کوتاهی دارند، کارایی بالا دارند. به‌منظور کاهش این محدودیت‌ها در این مقاله از یک دسته‌بند استفاده شده است، که از عامل‌های مختلفی به‌عنوان ویژگی استفاده می‌کند. روش پیشنهادی محدود به نوع خاصی از استخراج‌گرها نیست و جملات ورودی را به استخراج‌گر مناسب ارسال می‌کند. یک دسته‌بندی از روش‌های توضیح داده شده، به‌اختصار در جدول (۱) آورده شده است.

۳- جملات دشوار برای سامانه‌های استخراج آزاد اطلاعات

هدفِ تخمین‌زن دشواری جمله، شناسایی جملاتی است که استخراج آنها توسط سامانه‌های استخراج آزاد اطلاعات دشوار است. تعریف دشواری جمله در اینجا براساس دو فرضیه است:

۱) دشواری وابسته به سامانه است. بدین معنی که یک جمله ممکن است برای استخراج توسط یک سامانه استخراج‌گر دشوار باشد و برای استخراج توسط سامانهٔ دیگر آسان/ساده باشد.

۲) دشواری از کیفیت استخراج ضعیف نمایان می‌شود. به‌عنوان نمونه چندین جمله در جدول (۲) آورده شده است. این جدول خروجی دو سامانهٔ استخراج آزاد اطلاعات (ریورب و اگزمپلر) و نیز خروجی مطلوب را که به‌صورت دستی توسط انسان صورت گرفته نشان می‌دهد. جدول (۲) تغییرات دشواری جملات را برای این دو سامانهٔ مختلف نشان می‌دهد.

مطابق [30] درصورتی‌که یک مجموعه از موجودیت‌های اسمی E و یک مجموعه از نمونه روابط $R=\{r_1, \ldots, r_m\}$ وجود داشته باشد، یک نمونه رابطه دودویی، رکوردی به‌صورت $r_i = (a_1, p, a_2)$ است که در آن p یک گزاره و a_i یک آرگومان است و یک آرگومان می‌تواند یک موجودیت یا یک نمونه رابطه باشد ($a_j \in EUR$). نقش یک آرگومان توسط تابع $\rho(r_i, a_j) \rightarrow \{subject, direct\ object, prep\ object\}$ تعریف می‌شود که در آن مفعول حرف اضافه می‌تواند نقش‌های زیادی داشته باشد (برای هر حرف اضافه در زبان یک نقش می‌تواند داشته باشد).

(جدول-۱): مرور کلی بر روش‌های استخراج آزاد اطلاعات

(Table-1): A review of open information extraction methods

	نام روش	سال	ورودی مدل	مدل استخراج (روش تولید)
روش‌های سطحی	TextRunner[5]	۲۰۰۸	قواعد مکاشفه‌ای، ویژگی‌های سطح بالا	یادگیری خودناظر (دسته‌بند بیزین میدان تصادفی شرطی)
	Reverb[6]	۲۰۱۱	قیدهای واژگانی و نحوی، واژه‌نامه‌روابط	استفاده از عبارت منظم، مبتنی بر قواعد
	SONEX[15]	۲۰۱۲	بردار ویژگی‌ها از محتوای متن بین موجودیت‌های اسمی	الگوریتم HAC و Buckshot
	WOEPos[7]	۲۰۱۰	داده‌های ساخت‌یافته (ویکی پدیا)	یادگیری نیمه‌نظارتی (نظارت دور) و دسته‌بند میدان تصادفی شرطی
	R2A2[6]	۲۰۱۱	قیدهای واژگانی و نحوی، واژه‌نامه‌ی روابط	استفاده از عبارت منظم و رگرسیون منطقی
روش‌های عمیق	ZORE [16]	۲۰۱۴	برچسب‌زنی اجزای کلام و تجزیه درخت‌های تجزیهٔ وابستگی	الگوریتم انتشار
	Dep-OE [10]	۲۰۱۲	تجزیهٔ وابستگی	بدون یادگیری، مبتنی بر قواعد
	LSOE[19]	۲۰۱۳	متونی با برچسب اجزای کلام	بر مبنای الگوهای لغوی-نحوی
	KrakeN [9]	۲۰۱۳	قواعد مکاشفه‌ای و تجزیه وابستگی	بدون یادگیری، مبتنی بر قواعد
	OLLIE[11]	۲۰۱۲	برچسب‌زنی اجزای کلام و تجزیه درخت‌های تجزیه وابستگی	یادگیری الگوها از درخت وابستگی
	Patty [22]	۲۰۱۲	تجزیهٔ وابستگی	بدون یادگیری، مبتنی بر قواعد مکاشفه‌ای
	ClauseIE [4]	۲۰۱۳	تجزیهٔ وابستگی	بدون یادگیری، مبتنی بر دانش زبانی در دستور زبان انگلیسی
	Exemplar[13]	۲۰۱۳	تجزیهٔ وابستگی	قواعد دست‌نویس روی درخت وابستگی
	CSD-IE [23]	۲۰۱۳	جملات تجزیهٔ سازه‌ای شده	از تجزیه‌ی جملات و اعمال قواعد استنباطی
	R-OpenIE[25]	۲۰۱۶	برچسب‌زنی اجزای کلام	مبتنی بر قاعده با کمک مبدل وضعیت-متناهی آبشاری
	TreeKernel[26]	۲۰۱۳	تجزیهٔ وابستگی	الگوهای نحوی و نیز ماشین بردار پشتیبانی با کرنل‌های درخت وابستگی
روش‌های ترکیبی	EFFICIENS[30]	۲۰۱۵	متونی با برچسب اجزای کلام	تجزیه وابستگی و برچسب‌زن نقش معنایی
	TR-DOE[1]	۲۰۱۶	متون با برچسب اجزای کلام و تجزیه درخت‌های تجزیه وابستگی	ترکیب با کمک طول جمله و ضریب اطمینان
	RV-DOE[1]	۲۰۱۶	متون با برچسب اجزای کلام و تجزیه درخت‌های تجزیه وابستگی	ترکیب با کمک طول جمله و ضریب اطمینان
	RECALLHYBRID [27]	۲۰۱۱	برچسب‌زنی نقش معنایی	بدون یادگیری، مبتنی بر قواعد دستی
	PRECHYBRID[27]	۲۰۱۱	برچسب‌زنی نقش معنایی	بدون یادگیری، مبتنی بر قواعد دستی

سطرهای جدول (۲) بر اساس این تعریف استخراج شده‌اند. کلمات اختیاری که می‌توانند در نمونه رابطه قرار گیرند، توسط آکولاد مشخص شده‌اند. با داشتن یک جمله هر سامانه باید نمونه روابط شرح‌داده‌شده را در آن جمله استخراج کند. برای هر جمله در جدول (۲) ساختاری نحوی در آن وجود دارد که عمل استخراج را دشوار می‌سازد. کمبود این دانش در ساختار ریورب منجر به استخراج ضعیف در خروجی آن می‌شود. در مقابل، استخراج برای اگزمپلر ساده است؛ زیرا از تعدادی ابزار در ساختارش استفاده می‌کند تا چنین ساختارهای پیچیده‌ای را به‌سادگی استخراج کند. دلایل متعددی وجود دارد که استخراج رابطه را دشوار می‌سازد. بیشتر این دلایل به عمقی که سامانهٔ استخراج آزاد اطلاعات در آن کار می‌کند، مرتبط است.

از آنجایی که دشواری را براساس کیفیت استخراج مدل‌سازی می کنیم، مسائلی را در نظر می‌گیریم که در ساختار سامانه منعکس می شوند. در اینجا دشواری جمله از چشم‌انداز سامانه های استخراج آزاد اطلاعات در نظر گرفته شده است؛ جملات دشوار، جملاتی هستند که استخراج آنها توسط سامانه های استخراج آزاد اطلاعات سطحی دشوار است و روابط ناقص یا نادرستی را تولید می کند. شناسایی نمونه روابط از این جملات نیازمند به ابزارهای عمیق مانند برچسب زن نقش معنایی است. مطابق نتایج موجود در [13] برخی روش‌های سطحی نسبت به روش‌های عمیق، ده برابر جملات بیشتری را در مدت زمان یکسان مدیریت می‌کنند. دشواری جمله در چندین کاربرد از پردازش زبان طبیعی مانند اندازه‌گیری دشواری ترجمه [31]، ارزیابی قابلیت اطمینان تجزیه‌گرها [32]، اندازه‌گیری دشواری متن [33] و خوانایی متن [34] و غیره کاربرد دارد. ایدهٔ این کار می‌تواند در وظایف دیگری از پردازش زبان طبیعی نظیر پردازش گفتار، پرسش و پاسخ و موتورهای جستجو وارد شود و آنها نیز از تشخیص خودکار زیروظایف دشوار بهره‌مند شوند.

(جدول-۲): جملات دشوار و ساده نمونه

(Table-2): Sample Difficult and easy sentences

جمله	Daniel Jurafsky was recognized by the MacArthur.	Barak Obama is the president of the United States.	Dr. Joan Clos, the executive director of the United Nations Human Settlements Programme said urban planning should be democratic.
برچسب انسانی	(MacArthur, recognized, Daniel Jurafsky)	(Barak Obama, is {the} president {of the}, United States)	({Dr.} Joan Clos, executive director {of}, {the} United Nations Human Settlements Programme) ({Dr.} Joan Clos, said, (urban planning, should be, democratic))
خروجی ریورب	-	(Barak Obama, is, the president of the United States)	(the executive director of the United Nations Human Settlements Programme, said, urban planning)
خروجی اگزمپلر	(MacArthur, recognized, Daniel Jurafsky)	(Barak Obama, is president, United States)	(Dr. Joan Clos, the executive director, United Nations Human Settlements Programme)

که عمل استخراج را دشوار می‌سازد. کمبود این دانش در ساختار ریورب منجر به استخراج ضعیف در خروجی آن می‌شود. در مقابل استخراج برای اگزمپلر ساده است؛ زیرا از تعدادی ابزار در ساختارش استفاده می‌کند تا چنین ساختارهای پیچیده‌ای را به‌سادگی استخراج کند. دلایل متعددی وجود دارد که استخراج رابطه را دشوار می‌سازد. بیشتر این دلایل به عمقی که سامانه استخراج آزاد اطلاعات در آن کار می‌کند مرتبط است. از آنجایی که دشواری را براساس کیفیت استخراج مدل‌سازی می‌کنیم، مسائلی را در نظر می‌گیریم که در ساختار سامانه منعکس می‌شوند. در اینجا دشواری جمله از چشم‌انداز سامانه‌های استخراج آزاد اطلاعات در نظر گرفته شده است، جملات دشوار جملاتی هستند که استخراج آنها توسط سامانه‌های استخراج آزاد اطلاعات سطحی دشوار است و روابط ناقص یا نادرستی را تولید می‌کند. شناسایی نمونه روابط از این جملات نیازمند به ابزارهای عمیق مانند برچسب‌زن نقش معنایی است. مطابق نتایج موجود در [13] برخی روش‌های سطحی نسبت به روش‌های عمیق ده برابر جملات بیشتری را در مدت‌زمان یکسان مدیریت می‌کنند. دشواری جمله در چندین کاربرد از پردازش زبان طبیعی مانند اندازه‌گیری دشواری ترجمه [31]، ارزیابی قابلیت اطمینان تجزیه‌گرها [32]، اندازه‌گیری دشواری متن [33] و خوانایی متن [34] و غیره کاربرد دارد. ایدهٔ این کار می‌تواند در وظایف دیگری از پردازش زبان طبیعی نظیر پردازش گفتار، پرسش و پاسخ و موتورهای جستجو وارد شود و آنها نیز از تشخیص خودکار زیروظایف دشوار بهره‌مند شوند.

۴- روش پیشنهادی برای استخراج آزاد اطلاعات

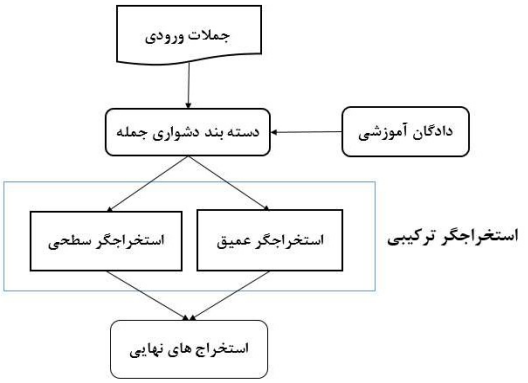
دو نوع تحلیل زبانی (سطحی و عمیق) برای استخراج رابطه در سامانه‌های استخراج آزاد اطلاعات مطرح است. از آنجایی که تمرکز استخراج آزاد اطلاعات به پیداکردن همه حقایق مفید از مجموعه بزرگ و ناهمگنی مانند وب در زمان معقول

است، روش‌هایی که از ابزارهای عمیق پردازش زبان طبیعی (ابزارهای تحلیل نحوی و معنایی مانند برچسب‌گذاری نقش معنایی و تجزیه‌گر) استفاده می‌کنند، پرهزینه بوده و مقیاس‌پذیری کمتری دارند. علاوه‌براین تنها استفاده از این ابزارها که برای تعداد محدودی از زبان‌های طبیعی فراهم بوده و نتایج نه‌چندان مطلوبی را تولید می‌کنند، معقول به‌نظر نمی‌رسد. از طرفی تحلیل دستی عمیق نیز کاری دشوار، زمان‌بر و پر هزینه است. روش دیگر برای استخراج روابط تکیه بر فقط تحلیل زبانی سطحی با استفاده از تجزیه‌گر سطحی، برچسب‌گذار اجزای کلام، لم‌یاب و... است. ابزارهای خودکار برای تحلیل سطحی برای تعداد زیادی از زبان‌ها موجود و به اندازه کافی قابل اعتماد هستند. این استخراج‌گرها سریع اما محدود به تحلیل نحوی سطحی هستند که رسیدن به بیشینه کارایی را محدود می‌کنند.

درواقع نیاز به سامانه‌ای است که استفاده مؤثر از زمان موجود را قادر می‌سازد و یک توازن معقول بین دقت و بازخوانی را پیشنهاد دهد. با توجه به اینکه هر کدام از این روش‌ها نقاط قوت و ضعف مختص خودشان را دارند، یکی از اهداف این مقاله، توسعه یک روش ترکیبی با درنظرگرفتن مشخصه‌های مثبت هر کدام این رویکردها است. ما کاربرد مشخصه‌های هر دو نوع سطحی و عمیق را برای استخراج رابطه به‌منظور دستیابی به کارایی بالا بررسی کردیم. مزایای این دو نوع استخراج‌گر ما را بر آن داشت تا به توسعهٔ روشی بپردازیم که نقاط مثبت هر دو را داشته باشد. یک چارچوب برای استخراج اطلاعات ترکیبی با ترکیب قدرت ریورب و اگزمپلر پیشنهاد شده است. شکل (۱) چارچوب کلی این روش را نشان می‌دهد.

دسته‌بند دشواری و استخراج‌گر ترکیبی قسمت‌های اصلی روش پیشنهادی ما هستند. برای هر جمله، دسته‌بند دشواری مناسب‌ترین سامانه برای پردازش آن را پیدا می‌کند. همان‌طور که در شکل مشخص شده، استخراج‌گر ترکیبی شامل دو جز اصلی است، یک استخراج‌گر سطحی (ریورب) و یک استخراج‌گر عمیق (اگزمپلر). استخراج‌های نهایی با اجتماع خروجی این دو استخراج‌گر به‌وجود می‌آید. در ادامه هر یک از این دو مؤلفه شرح داده می‌شود.

دسته‌بند دشواری جمله: این مؤلفه به تخمین دشواری عمل استخراج رابطه برای جملات ورودی سامانهٔ استخراج آزاد اطلاعات تمرکز دارد. هدف این مؤلفه این است که جمله ورودی را به هدف افزایش کارایی، به دسته‌بند مناسب انتساب کند. تخمین‌زن دشواری برای استخراج آزاد اطلاعات هر جمله را به‌صورت متوالی می‌خواند. با داشتن یک جمله، دسته‌بندِ دشواری با کمک مجموعه‌ای از ویژگی‌های پیشنهادی دشواری جمله ورودی را برای عمل استخراج پیش‌بینی می‌کند.



(شکل-۱): چارچوب روش پیشنهادی: دسته‌بند، بهترین استخراج‌گر را به‌کار می‌گیرد.

(Figure-1): Framework of proposed approach: Difficulty classifier exploits the best of both the shallow and the deep OIE extractors

به عبارت دیگر برای هر جمله ورودی دسته‌بندِ دشواری، مناسب‌ترین سامانه را برای پردازش آن جمله انتخاب می‌کند. درصورتی‌که نمونهٔ ورودی برای پردازش با استخراج‌گر سطحی، دشوار در نظر گرفته شود، باید توسط یک استخراج‌گر عمیق پردازش شود. در این روش بهترین مسامحه بین کارایی و هزینهٔ محاسباتی به‌دست آمده است.

برای دسته‌بندی دودویی جملات، از دسته‌بند لجستیک منطقی استفاده شده است. به‌دلیل نتایج طبقه‌بندی قوی، این دسته‌بند برای مسائل دسته‌بندی مختلف در زبان‌شناسی محاسباتی استفاده می‌شوند. علاوه بر طبقه‌بندی، نیاز به پیداکردن معیاری داریم که بتوان شدت دشواری استخراج را تخمین زد. برای این منظور از امتیاز دسته‌بند به‌عنوان معیار دشواری استفاده می‌شود.

مسأله تخمین دشواری را به‌عنوان مسأله دسته‌بندی فرمول‌بندی کردیم که هدف انتساب برچسب طبقهٔ ساده یا دشوار به جملات نامزدی مانند s براساس دسته‌بندی مانند c است. با این کار جمله به استخراج‌گر مناسب انتساب می‌شود.

$$c : s \rightarrow \{easy, difficult\} \qquad (۱)$$

وظیفه دسته‌بند دشواری (دسته‌بند احتمالی رگرسیون لجستیک) انتخاب برچسب مناسب از برچسب‌های خروجی y که باید به ورودی x اختصاص داده شوند و نیز انتخاب مقداری از y است که شرط $P(x|y)$ را بیشینه کند. ویژگی f_i دارای مقادیر حقیقی اما در پردازش زبانی معمول‌تر این است که این

خصوصیات دارای مقادیر مبنای دو باشند. یک ویژگی که تنها دارای مقادیر صفر یا یک باشد یک تابع شاخص نام دارد. همچنین هر ویژگی تنها خصیصه‌ای از نمونهٔ x نیست؛ بلکه خصیصه‌ای از نمونه x و نیز طبقهٔ خروجی نامزد c است. بنابراین در بیشینه بی‌نظمی به‌جای نمایش با f_i یا $f_i(x)$ آنها را با $f_i(c,x)$ نشان می‌دهیم که به معنای ویژگی i برای طبقهٔ خاص c برای نمونه خاص x است [35]. با استفاده از عامل هنجارسازی Z و با مشخص‌کردن تعداد ویژگی‌ها به‌عنوان N می‌توان به معادله نهایی دست پیدا کرد که احتمال تعلق y به طبقهٔ c با داشتن x در بیشینه بی‌نظمی را محاسبه می‌کند:

$$P(c|x) = \frac{1}{Z}\exp\left(\sum_{i=1}^{N} w_i f_i(c,x)\right) \qquad (۲)$$

مسألهٔ مورد توجه دسته‌بند، این است که کدام استخراج‌گر باید هر جمله ورودی را پردازش کند تا تعداد نمونه‌های درست استخراج‌شده و درنتیجه کارایی به بیشینه برسد. این مدل می‌تواند ویژگی‌های پیشنهادی را بگیرد و احتمال دشواری استخراج نمونه خاص و پردازش آن توسط استخراج‌گر عمیق را برگرداند. درواقع بهترین مسامحه بین کارایی و هزینه محاسباتی توسط تنظیم پارامترهای ترکیب مؤثر به‌دست خواهد آمد. این دسته‌بند، استخراج‌گر عمیق را برای جملاتی از ورودی اعمال می‌کند که استخراج‌گر سطحی قادر به استخراج صحیح از آن نیست و در شرایطی که پردازش جمله با استخراج‌گر عمیق بهینه نباشد، استخراج‌گر سطحی استفاده می‌شود؛ بدین ترتیب از منابع استفادهٔ مؤثرتری می‌شود.

برای این منظور از دسته‌بندهای احتمالاتی رگرسیون لجستیک استفاده شده است که به‌طور خودکار یک امتیاز دشواری برای هر جمله ورودی انتساب شود. رگرسیون لجستیک گاهی به‌عنوان مدل‌سازی بی‌نظمی بیشینه و یا MaxEnt نیز شناخته می‌شود. رگرسیون لجستیک متعلق به خانواده دسته‌بندهای معروف به نام دسته‌بندهای توانی و یا لگاریتمی–خطی است. مانند روش بیز، این الگوریتم از طریق استخراج یک مجموعه از خصوصیات وزن‌دار از ورودی، لگاریتم‌گرفتن از آنها و ترکیب آنها به‌صورت خطی عمل می‌کند (بدین معنی که هر ویژگی به‌صورت ضرب در وزن آن و سپس جمع‌کردن نتایج است). درواقع رگرسیون لجستیک به دسته‌بندی‌هایی اشاره دارد که یک مشاهده را به دو طبقه تفکیک می‌کند و رگرسیون لجستیک چندجمله‌ای، زمانی استفاده می‌شود که بیش از دو طبقه وجود داشته باشد [36,37]. از ابزار وکا[1] برای پیاده‌سازی رگرسیون لجستیک استفاده شده است. روش یادگیری وزن‌ها گرادیان کاهشی است.

ویژگی‌های عمیق نسبت به ویژگی‌های سطحی می‌تواند دقت و بازخوانی را بهبود دهد؛ اما سرعت استخراج افزایش می‌یابد. برای نمونه، ویژگی‌های مبتنی بر تجزهٔ وابستگی می‌تواند به مدیریت روابط پیچیده و با فاصلهٔ دور در جملات دشوار کمک کند. چنین شرایطی به‌طورمعمول نمی‌تواند توسط ویژگی‌های سطحی شناسایی شود. با درنظرگرفتن هزینهٔ محاسباتی مربوط به ویژگی‌های غنی، در اینجا از ۶۱ ویژگی سبک‌وزن استفاده شده است. تمام این ویژگی‌ها مقیاس‌پذیر و مستقل از دامنه هستند و می‌توانند در زمان استخراج بدون استفاده از ابزارهای عمیق ارزیابی شوند. این ویژگی‌ها می‌توانند از استخراج‌گرهای زیرین نیز به‌وجود آیند.

ویژگی‌های مختلفی برای دسته‌بند دشواری در نظر گرفته شده که به‌طورکلی از خصوصیات ساختار نحوی جملات و برخی ویژگی‌های معنایی آن‌ها تشکیل شده است. برای محاسبهٔ مقادیر ویژگی‌های نحوی نیاز به استفاده از برچسب‌زن اجزای کلام وجود دارد. برخی از ویژگی‌های نحوی ممکن است از ترتیب کلمات به‌وجود آیند. این ویژگی‌ها به دسته‌بند اجازه می‌دهند تا میزان سختی را تخمین بزند که سامانه هنگام استخراج نمونه‌ها از جمله با آن مواجه می‌شود. نمونه‌هایی از ویژگی‌های به‌کاررفته شامل طول جمله، وجود دنباله‌ای از برچسب‌های اجزای کلام درجمله، تعداد ایست‌واژه‌ها در جمله، وجود یا عدم وجود افعال شناخت و ارتباطی در جمله، عبارات منظم (برای شناسایی نقل قول، حرف بزرگ، کلمات پرسشی و غیره) است که در زیر تعدادی از آن‌ها بیان شده است:

- جمله شامل دست‌کم یک فعل است که دو موجودیت اسمی در طرفین آن قرار دارند. مانند:

A suicide car bomber **attacked** Iraq's largest newspaper.

- جمله شامل دست‌کم یک فعل است که دو موجودیت اسمی در طرفین آن قرار دارند؛ به‌طوری‌که با نخستین موجودیت اسمی شروع می‌شود؛ مانند:

Calcium prevents osteoporosis.

- جمله شامل ضمایر who، which، that است؛ مانند:

Chicago, **which** is located in Illinois has three million residents.

[1] https://www.cs.waikato.ac.nz/ml/weka/

- جمله شامل دستِ‌کم یک فعل است که دو موجودیت اسمی در طرفین آن قرار دارند؛ به‌طوری‌که نخستین موجودیت اسمی، اسم خاص است؛ مانند:

Barack Obama was elected as president.

- جمله شامل دستِ‌کم یک فعل است که دو موجودیت اسمی در طرفین آن قرار دارند؛ به‌طوری‌که دومین موجودیت اسمی، اسم خاص است؛ مانند:

Google acquired **YouTube** in 2006.

- جمله شامل دستِ‌کم یک فعل است که دو موجودیت اسمی در طرفین آن قرار دارند به‌طوری‌که if قبل از نخستین موجودیت وجود دارد؛ مانند:

If Trump wins the election, the House and the Senate will
definitely be in Republican hands

- طول جمله بزرگتر از ۱۰ است؛ مانند:

John McCain fought hard against Barak Obama, but finally lost
the election.

- جمله شامل دستِ‌کم دو موجودیت اسمی است و بعد از دومین موجودیت اسمی فعل وجود دارد؛ مانند:

After winning the Superbowl, the Saints **are** now the top dogs of the NFL.

- جمله شامل افعال ارتباطاتی[1] است؛ مانند:

Early astronomers **believed** that the earth is the of the universe.

- تعداد ضمایر نسبی[2] در جمله یک یا بیشتر است؛ مانند:

A federal judge said **that** Trump does not have the right to block
people from following his Twitter posts.

این مجموعه از بردارهای ویژگی برچسب‌زده‌شدهٔ خودکار به‌عنوان ورودی به دسته‌بند رگرسیون لجستیک به‌کار می‌رود. هر جمله براساس خروجی دسته‌بند به یک استخراج‌گر ارسال می‌شود.

استخراج‌گر ترکیبی: یک مؤلفه ترکیبی مبتنی بر استخراج‌گرهای سطحی و عمیق است که به استخراج بهینه روابط از جملات زبان طبیعی می‌پردازد. استخراج‌گر ترکیبی با به‌کارگیری استخراج‌گرهای عمیق برای جملاتی که استخراج‌گرهای سطحی قادر به استخراج صحیح از آنها نیستند و یا پردازش آنها توسط استخراج‌گرهای عمیق بهینه نمی‌باشد، قادر به تخصیص بهتر منابع محاسباتی و حداقل استفاده از ابزارهای عمیقِ پردازش زبان طبیعی و اجتناب از اتلاف این منابع در جملاتی است که احتمال بهبود کارایی در آنها کم است. این مؤلفه از زمان موجود، استفادهٔ مؤثر می‌کند و مقیاس‌پذیری را افزایش می‌دهد.

[1] communication
[2] relative pronouns

در این مؤلفه ترکیبی از ریورب به‌عنوان استخراج‌گر سطحی و از اگزمپلر به‌عنوان استخراج‌گر عمیق استفاده شده است. یک مقایسه آزمایشگاهی هدفمند و منصفانه‌ای از ده رویکرد اخیر در [13] و [30] انجام شده است. مطابق این پژوهش، ریورب سریع‌ترین روش سطحی است که براساس تطبیق الگوها روی برچسب‌های اجزای کلام کار می‌کند. ریورب از ویژگی‌های نحوی سطحی برای تولید روابط معنایی استفاده و درنتیجه مقیاس‌پذیری با اندازهٔ پیکره را تضمین می‌کند.

SONEX نیز یک استخراج‌گر آزاد سطحی است که نتایج قابل مقایسه‌ای با ریورب دارد و برای غلبه بر چالش‌های اصلی به‌کارگیری سامانه‌های استخراج آزاد اطلاعات در وبلاگ‌ها طراحی شده است که از الگوریتم خوشه‌بندی برای گروه‌بندی جفت‌ها با بافتار متنی مشابه با همدیگر در مقیاس بزرگ به‌کار می‌رود. علاوه‌بر چالش‌های مربوط به خوشه‌بندی در مقیاس بزرگ (زمان و فضا)، این روش روابط را در سطح پیکره شناسایی می‌کند. از آنجایی که روش پیشنهادی مبتنی بر جمله است (بدین معنی که فرایند استخراج می‌تواند به‌تنهایی از یک جمله صورت گیرد.) از ریورب به‌عنوان استخراج‌گر سطحی تشکیل‌دهندهٔ مؤلفهٔ استخراج‌گر ترکیبی استفاده شده است. اگزمپلر نیز استخراج‌گر عمیق جدیدی است که در مقایسه با سایر استخراج‌گرهای عمیق، بالاترین کارایی را دارد و از طرفی در مقایسه با دیگر استخراج‌گرهای عمیق مانند استخراج‌گرهای مبتنی بر برچسب‌زن نقش معنایی سریع‌تر است و زمان اجرای کمی دارد. این استخراج‌گر براساس اعمال قواعد روی درخت‌های تجزیهٔ وابستگی کار می‌کند و در مقایسه با ریورب به‌خوبی عمل کرده و کارایی بالایی دارد. توضیحات بیشتر از ساختار این دو استخراج‌گر آزاد اطلاعات در بخش ۲ آورده شده است.

سامانه‌های تشکیل‌دهندهٔ هستهٔ استخراج‌گر بر پایهٔ ابزارهای تحلیل زبانی سطحی و عمیق هستند و نیاز به دادهٔ آموزشی ندارند. این روش مستقل از سامانه‌های تشکیل‌دهندهٔ آن است و می‌توانند با دیگر سامانه‌های سطحی یا عمیق نیز طراحی شوند.

ایدهٔ اولیه‌ای از این روش براساس دو پارامتر ترکیب: طول جمله و ضریب اطمینان در [1] نشان داده شده است. درحالی‌که در ساختار روش‌های مشابهٔ قبلی، ابزار تحلیل زبانی سطحی یا عمیق به‌تنهایی دستِ‌کم یکبار برای تمام جملات ورودی به‌کار گرفته می‌شود. تا جایی‌که می‌دانیم، نخستین‌بار است که یک روش برای تفکیک ورودی به یک استخراج‌گر

مناسب، به‌منظور دستیابی به کارایی بالا طراحی شده است. روش پیشنهادی به‌ویژه زمانی مفید است که مجموعهٔ دادهٔ بزرگی داشته باشیم و زمان پردازش محدود باشد. در این شرایط، استخراج‌گر ترکیبی ما از زمان موجود، استفادهٔ مؤثری و بهترین الگوریتم را بر اساس زمان محاسباتی موجود اجرا می‌کند. علاوه‌براین گاهی فقط تعداد جملات کمی، درکل مجموعهٔ داده نمونه‌های بهتری را با ابزارهای پردازش زبان طبیعی عمیق تولید می‌کنند. از طرف دیگر، در این شرایط استخراج‌گری که از ابزارهای پردازش زبان طبیعی عمیق برای کل ورودی‌ها بهره می‌گیرد، منابع محاسباتی را برای سایر جملات در مجموعه داده به هدر می‌دهد. زمانی که هر دو استخراج‌گر استخراج صحیحی را تولید می‌کنند، روش پیشنهادی، استخراج‌گر سطحی را ترجیح خواهد داد و بنابراین کارایی بهبود می‌یابد.

۵- نتایج آزمایش‌ها

تأثیر به‌کاربردن دسته‌بند دشواری در انتخاب مؤثر استخراج‌گرها ارزیابی و رفتار استخراج‌گرهای سطحی و عمیق بررسی شده است. یک دسته‌بند رگرسیون لجستیک آموزش داده شده است که یک جمله را به‌عنوان ورودی می‌گیرد و تصمیم می‌گیرد که آن جمله برای استخراج آزاد اطلاعات دشوار است یا آسان. داده‌های استاندارد طلایی و یک مجموعه از ویژگی‌ها برای آموزش دسته‌بند دشواری مورد نیاز است. کمبود مجموعهٔ دادهٔ استاندارد یکی از چالش‌های اصلی در ارزیابی سامانه‌های استخراج آزاد اطلاعات است. روش‌های ارزیابی موجود به ارزیابی دستی تکیه می‌کنند [4, 7, 9, 10, 12, 21, 23, 26, 38] که محدودیت اصلی آنها این است که مقیاس‌پذیر نیستند. ترکیب مجموعه داده‌های موجود و ایجاد مجموعه دادهٔ بزرگ‌تر نیز از سوی دیگر دشواری‌هایی دارد. تفاوت‌ها در برچسب‌زنی و روش ارزیابی، برخی از این چالش‌ها است. ایجاد دستی مجموعهٔ داده بزرگ نیز از سوی دیگر بسیار زمان‌بر و پرهزینه است. از آن‌جایی‌که کارایی در روش پیشنهادی بسیار حیاتی است، برچسب‌زدن خودکار نیز ممکن است، منجر به دقت و بازخوانی به‌نسبه پایین شود.

براساس منابع موجود، ما از مجموعه دادهٔ جدید [30] که به‌صورت دستی برچسب‌زده شده است، استفاده کردیم. این مجموعه داده نسبت به دیگر مجموعه داده‌هایی نظیر [4, 7, 9, 10, 12, 21, 23, 26, 38] بسیار بزرگ‌تر است. این مجموعه داده سعی می‌کند تا مشکلات مربوط به کمبود مقادیر داده‌های واقعی و نیز تفاوت‌ها در روش‌های ارزیابی را با فراهم‌کردن برچسب‌های قابل استفادهٔ مجدد که انعطاف‌پذیر هستند و می‌توانند برای ارزیابی محدودهٔ وسیعی از روش‌ها به‌کار روند، کاهش دهد [30].

این مجموعه داده جملاتی را از نیویورک‌تایمز (NYT-500)، پن تری بانک (PENN-100) و یک کرپوس عام برای وب (WEB-500) پوشش می‌دهد. WEB-500 شامل پانصد جملهٔ استخراج‌شده از قسمت‌های کوچکی از موتورهای جستجو است. این جملات اغلب ناقص بوده و از نظر گرامری نادرست هستند و چالش‌های متون وب را به‌خوبی نشان می‌دهند. NYT-500 نیز شامل پانصد جمله است که در آن جملات منفردی را از متون رسمی و متونی از پیکرهٔ نیویورک تایمز که به‌خوبی نوشته شده‌اند، نشان می‌دهد PENN-100 شامل یکصد جمله از پن‌تری بانک است که درهمین‌اواخر برای ارزیابی روش هسته درخت [26] به‌کار رفته است. NYT-500 و WEB-500 به‌عنوان مجموعهٔ دادهٔ آموزشی و PENN-100 به‌عنوان مجموعه دادهٔ آزمایش به‌کار رفته است.

داده‌های طلایی شامل مجموعه‌ای از جملاتی است که برچسب دشوار یا آسان دارند. این مجموعه داده به‌صورت دستی و برحسب خروجی‌های ریورب و اگزمپلر برچسب زده شده است. با داشتن یک پیکره، روش پیشنهادی باید به‌گونه‌ای جملات را برای پردازش به‌وسیلهٔ استخراج‌گرهای سطحی/عمیق انتخاب کند که تعداد نمونه‌های درست استخراج‌شده به بیشینه برسد. به‌عبارت دیگر آن استخراج‌گری را انتخاب می‌کند که خروجی درستی را تولید کند، زمانی که دیگری خروجی نادرستی را تولید می‌کند. یک جمله برچسب آسان زده می‌شود، اگر استخراج‌گر سطحی نتیجهٔ درستی را تولید کند. در شرایطی که استخراج‌گر سطحی نتیجهٔ نادرستی را تولید کند، برچسب دشوار زده می‌شود؛ به‌جز در حالتی‌که استخراج‌گر عمیق نیز نتایج نادرستی را تولید کند. یک جمله همچنین برچسب آسان زده می‌شود، اگر استخراج‌گر سطحی هیچ خروجی برای آن جمله نداشته باشد؛ ولی استخراج‌گر عمیق نتایج نادرستی را تولید کند. در این شرایط اگر استخراج‌گر عمیق، نتایج درستی را تولید کند، جمله به‌صورت دشوار برچسب زده خواهد شد.

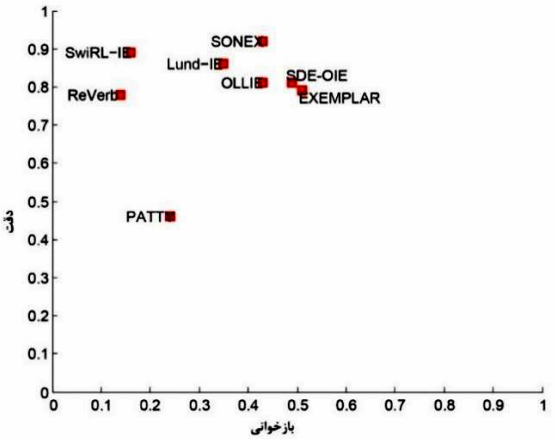
ارزیابی ما روی استخراج نمونه‌های رابطه در سطح جمله تمرکز دارد. معیارهای به‌کار رفته در ارزیابی شامل دقت، بازخوانی و معیار-f است. دقت به‌صورت نسبت تعداد نمونه‌های درست استخراج‌شده به تعداد کل نمونه‌های استخراج‌شده تعریف می‌شود. بازخوانی نیز نسبت تعداد نمونه‌های درست استخراج‌شده به تعداد کل نمونه‌های درست

تعریف می‌شود. معیار-f نیز متوسط دقت و بازخوانی را نشان می‌دهد [1].

نمونه‌های رابطه با مقادیر امتیازهای مساوی یا بزرگ‌تر از یک حدآستانهٔ خاص متعلق به طبقهٔ یک و نمونه‌های با مقادیر امتیازهای کمتر از آن حد آستانه متعلق به طبقهٔ صفر در نظر گرفته شده‌اند. مقادیر مختلف امتیازهای رگرسیون لجستیک بررسی و مشاهده شد که حد امتیاز ۰/۶ بالاترین دقت را به‌وجود می‌آورد.

نمودار شکل (۲) دقت و بازخوانی هر یک از روش‌ها را نشان می‌دهد. یک تفاوت جزیی بین دقت روش پیشنهادی و استخراج‌گرهای تشکیل‌دهندهٔ آن وجود دارد. ریورب و اگزمپلر دقت به‌نسبه بالایی را به‌دلیل طراحی الگوهای به‌نسبه مناسب برای استخراج رابطه دارند.

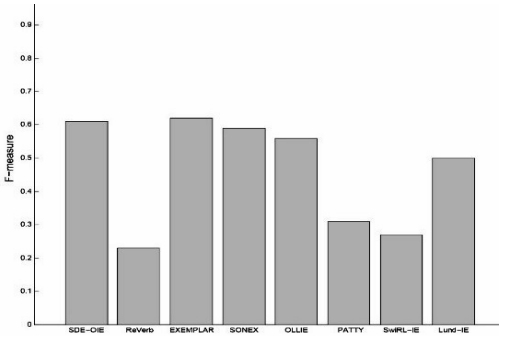
دقت روش پیشنهادی بالاتر از دقت ریورب و اگزمپلر است. این امر می‌تواند به‌دلیل دقت بالای استخراج‌گرهای تشکیل‌دهندهٔ آن (یعنی ریورب و اگزمپلر) به‌وجود آید. از نظر دقت، سانکس از دیگر روش‌ها بهتر است از آنجایی‌که طراحی مبتنی بر الگو آن قادر به شناسایی صحیح گزاره‌های به‌وجود آمده توسط اسم است. اگزمپلر بالاترین بازخوانی را نسبت به روش‌های دیگر دارد. روش پیشنهادی به‌طورتقریبی سطح یکسانی از بازخوانی را با اگزمپلر و مقدار بسیار بالاتری نسبت به ریورب دارد. بازخوانی اگزمپلر به‌دلیل اینکه می‌تواند نمونه‌های درست بیشتری را شناسایی کند زیاد است به‌ویژه آنهایی که شامل گزاره‌هایی از فعل+اسم هستند.



(شکل-۲): روش پیشنهادی بالاترین دقت را نسبت به استخراج‌گرهای تشکیل‌دهنده‌اش و مقدار بازخوانی چشم‌گیری را نسبت به استخراج‌گر سطحی تشکیل‌دهنده‌اش دارد.

(Figure-2): Proposed approach achieves higher precision than its underlying Open IE systems, and higher recall than ReVerb.

شکل (۳) نمودار معیار-f را برای هر کدام از روش‌ها نشان می‌دهد. اگزمپلر بالاترین مقدار را در مقایسه با سایر روش‌ها دارد. این به‌طور عمده به این دلیل بازخوانی به‌نسبه بالای اگزمپلر در مقایسه با دیگر روش‌ها است. روش پیشنهادی و اگزمپلر در سطح بسیار نزدیکی از معیار-f قرار دارند. این نشان می‌دهد که روش پیشنهادی دستِ‌کم به‌خوبی استخراج‌گر عمیق تشکیل‌دهنده‌اش است. ریورب، بازخوانی پایین‌تری از دیگر روش‌های دارد؛ این امر به‌دلیل ضعف ذاتی ابزارهای سطحی در شناسایی نمونه‌های روابط است که منجر به افت شدید در مقدار معیار-f آن می‌شود.



(شکل-۳): روش پیشنهادی و EXEMPLAR تقریباً مقدار معیار-f یکسانی دارند. مقدار معیار-f آنها بالاتر از سایر روش‌ها است.

(Figure-3) :Proposed method and EXEMPLAR have almost the same F-measure. Their F-measure is better than the others.

نتایج آزمایش‌ها اثبات می‌کند که تلفیق مناسبی از استخراج‌گرهای سطحی و عمیق، تعداد خروجی‌های نادرست را کاهش داده و تعداد خروجی‌های درست را افزایش می‌دهد و درنتیجه منجر به افزایش کارایی می‌شود. روش ترکیبی پیشنهادی قادر به پوشش ضعف‌های استخراج‌گر سطحی است و منجر به افزایش چشم‌گیری در کارایی آن می‌شود و نیز به مقدار معیار-f ای می‌رسد که به‌طورتقریبی سه برابر ریورب است.

زمان محاسباتی برای انواع مختلف استخراج‌گرها تغییر می‌کند. تمایز صریحی به اندازهٔ به‌طورتقریبی یک درجه بزرگی بین روش‌هایی وجود دارد که بر پایهٔ تجزیهٔ معنایی (SwiRL-IE یا Lund-IE)، تجزیهٔ وابستگی (OLLIE، اگزمپلر و PATTY) و تجزیهٔ سطحی (ریورب و SONEX) هستند. ریورب از آنجایی‌که از الگوهای سطحی استفاده می‌کند سریع‌ترین روش است و بر هیچ ابزار عمیقی تکیه ندارد. همان‌طور که نتایج نشان می‌دهد، استخراج‌گرهای عمیق به‌طورمعمول هزینهٔ محاسباتی بالایی دارند. درکل، هر چه استخراج‌گر عمیق‌تر باشد، زمان محاسباتی‌اش بیشتر است.

در مدت زمانی یکسان، استخراج‌گرهای سطحی چندین برابر جملات بیشتری را نسبت به استخراج‌گرهای تجزیهٔ وابستگی پردازش می‌کنند و آنها نیز چندین برابر جملات بیشتری را نسبت به استخراج‌گرهای تجزیهٔ معنایی پردازش می‌کنند. روش پیشنهادی از زمان موجود، استفادهٔ مؤثری می‌کند و به موازنهٔ معقولی از دقت و بازخوانی دست می‌یابد. روش پیشنهادی به‌طورتقریبی مقدار یکسانی از معیار-f را همانند اگزمپلر دارد، اما در زمان پردازشی بسیار کمتر. این موضوع در مقیاس‌های بزرگی مانند داده‌های وب اهمیت زیادی پیدا می‌کند. وقتی تعداد جملات پردازش‌شده توسط ریورب زیاد است، زمان کل به زمان پردازش ریورب کاهش می‌یابد. یک نتیجهٔ جالب این است که با وجود داشتن دقت بالا، روش‌های مبتنی بر تجزیهٔ وابستگی (SwiRL-IE و Lund-IE) معیار-f پایین‌تری نسبت به روش پیشنهادی دارند و از طرفی نیاز به زمان محاسباتی بسیار بالایی دارند. توزیع جملات با برچسب‌های آسان و دشوار به‌ترتیب ۳۹٪ و ۶۱٪ است. دقت دسته‌بندی ٪۷۲/۲ و نرخ خطا ٪۲۷/۷ است. جدول (۵) ماتریس درهم‌ریختگی را برای دسته‌بند نشان می‌دهد. مشاهده می‌شود که خطای غالب زمانی اتفاق می‌افتد که یک جملهٔ دشوار به‌عنوان آسان طبقه‌بندی شود.

(جدول-۳): زمان محاسباتی(ثانیه) برای هر رویکرد
(Table-3): Computing time (per second) for each method.

نام رویکرد	مدت زمان(ثانیه)
روش پیشنهادی	0.38
ReVerb	0.02
EXEMPLAR	0.62
SONEX	0.04
OLLIE	0.14
PATTY	0.66
SwiRL-IE	2.17
Lund-IE	5.21

(جدول-۴): ماتریس درهم‌ریختگی مربوط به کارایی دسته‌بند
(Table-4): The confusion matrix for the performance of the difficulty classifier

داده‌های طلایی/داده‌های دسته‌بندی شده	آسان	دشوار
آسان	0.66	0.25
دشوار	0.33	0.75

۶- نتیجه‌گیری و کارهای آینده

هدف روش پیشنهادی این است که بتواند به دقت بالا در سرعت قابل قبول دست یابد؛ به‌طوری‌که قابلیت اجرا در مقیاس وب را داشته باشد. درواقع روش پیشنهادی به‌جای استفاده از فقط ابزارهای عمیق، سعی در استفاده از ویژگی‌های سطحی دارد تا با افزایش پیکرهٔ ورودی، مقیاس‌پذیر باشد. این مقاله روشی جدید را برای تخمین خودکار دشواری جملات در سامانه‌های استخراج آزاد رابطه نشان می‌دهد. از یک رگرسیون لجستیک با مجموعه‌ای از ویژگی‌های کارای مبتنی بر جمله برای تلفیق قدرت سامانهٔ استخراج آزاد اطلاعات سطحی با عمیق به‌کار گرفته شده است. دسته‌بند دشواری، جملات دشوار برای پردازش توسط استخراج‌گر عمیق را شناسایی می‌کند. بهترین مسامحهٔ بین زمان محاسباتی و معیار-f شناسایی می‌شود. نتایج آزمایش‌ها نشان می‌دهد که روش پیشنهادی به مقدار معیار-f بسیار بالاتر از استخراج‌گر سطحی تشکیل‌دهنده‌اش دست می‌یابد، همچنین این مقدار به‌طورتقریبی با استخراج‌گر عمیق تشکیل‌دهنده‌اش یکسان است؛ اما زمان پردازشی بسیار کمتر از آن دارد.

هدف استخراج آزاد اطلاعات این است که روش‌های استخراج اطلاعات به مقیاس و تنوعی به اندازهٔ دامنهٔ وب دست یابند. روش پیشنهادی فقط در صورتی جملات ورودی را به استخراج‌گر عمیق می‌فرستد که این کار لازم باشد. علاوه‌براین گاهی فقط جملات کمی در کل مجموعهٔ داده به کمک ابزارهای پردازش زبان طبیعیِ عمیق، نتایج بهتری تولید می‌کنند. در این شرایط روش پیشنهادی از زمان موجود استفادهٔ مؤثری می‌کند و بدین ترتیب مقیاس‌پذیری تقویت می‌شود. این روش قادر است تا منابع محاسباتی را بهتر تخصیص دهد و از اتلاف آنها جلوگیری کند و برای شرایطی مناسب است که زمان محاسباتی محدود و کارایی بالا مورد نیاز باشد.

این روش می‌تواند به‌خوبی به مسأله چند کلاسه تبدیل شود. برای این‌کار یک روش مبتنی بر برچسب‌گذاری نقش معنایی می‌تواند به‌عنوان عمیق‌ترین استخراج‌گر زیرین به‌کار رود. ویژگی‌های پیشنهادی برای محاسبه بسیار سریع هستند، که از جنبهٔ عملی بسیار مهم است. علاوه‌بر ویژگی‌های پیشنهادی ما، تعدادی ویژگی از استخراج‌گرهای زیرین نیز می‌تواند در دسته‌بند دشواری تلفیق شود. استفاده از ویژگی‌های معنایی نیز دانش زبانی عمیقی را به مدل وارد کرده اما هزینهٔ زمان در آنها زیاد است.

7- References ۷- مراجع

[1] V. Reshadat, M. Hoorali, and H. Faili, "A Hybrid Method for Open Information Extraction Based on Shallow and Deep Linguistic Analysis," *Interdisciplinary Information Sciences,* vol. 22, pp. 87-100, 2016.

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

[۳] نیما مولایی، حسین شیرازی. روش پیشنهادی برای استخراج اطلاعات مورد نیاز از متون نظامی. فصل‌نامه پردازش علائم و داده‌ها. ۱۳۹۱؛ ۹(۱): ۶۷-۸۰

[3] N. mollaei, A. Abdolahzadeh, H. A. Shirazi, *new approach to extract the required information from military documents*. JSDP. 2012; 9 (1): pp.67-80

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

[5] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld, "Open information extraction from the web," *Communications of the ACM,* vol. 51, pp. 68-74, 2008.

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

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

[8] A. Akbik and J. Broß, "Wanderlust: Extracting semantic relations from natural language text using dependency grammar patterns," in *WWW Workshop*, 2009.

[9] A. Akbik ,and A. 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*, 2012, pp. 52-56.

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

[11] V. Tablan, K. Bontcheva, D. Maynard, and H. Cunningham, "Ollie: on-line learning for information extraction," in *Proceedings of the HLT-NAACL 2003 workshop on Software engineering and architecture of language techno-logy systems-Volume 8*, 2003, pp. 17-24.

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

[13] F. Mesquita, J. Schmidek, and D. Barbosa, "Effectiveness and efficiency of open relation extraction," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing,* vol. 500, pp. 447–457, 2013.

[14] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open information extraction for the web," in *IJCAI*, 2007, pp. 2670-2676.

[15] Y. Merhav, F. Mesquita, D. Barbosa, W. G. Yee, and O. Frieder, "Extracting information networks from the blogosphere," *ACM Transactions on the Web (TWEB),* vol. 6, p. 11, 2012.

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

[17] Y.-H. Tseng, L.-H. Lee, S.-Y. Lin, B.-S. Liao, M.-J. Liu, H.-H. Chen, O. Etzioni, and A. Fader, "Chinese open relation extraction for knowledge acquisition," *EACL 2014,* p. 12, 2014.

[18] P. Gamallo and M. Garcia, "Multilingual open information extraction," in *Portuguese Conference on Artificial Intelligence*, 2015, pp. 711-722.

[19] C. Castella Xavier, S. de Lima, V. Lúcia, and M. Souza, "Open information extraction based on lexical-syntactic patterns," in *Intelligent Systems (BRACIS), 2013 Brazilian Conference on*, 2013, pp. 189-194.

[20] P. Cimiano ,and J. Wenderoth, "Automatically learning qualia structures from the web," in *Proceedings of the ACL-SIGLEX workshop on deep lexical acquisition*, 2005, pp. 28-37.

[21] M. Schmitz, R. Bart, S. Soderland, and O. Etzioni, "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*, 2012, pp. 523-534.

[22] N. Nakashole, G. Weikum, and F. Suchanek, "PATTY: a taxonomy of relational patterns with semantic types," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 1135-1145.

[23] H. Bast and E. Haussmann, "Open information extraction via contextual sentence decomposition," in *Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on*, 2013, pp. 154-159.

[24] H. Bast and E. Haussmann, "More informative open information extraction via simple inference," in *Advances in information retrieval*, ed: Springer, 2014, pp. 585-590.

[25] H. Lin, Y. Wang, P. Zhang, W. Wang, Y. Yue, and Z. Lin, "A Rule Based Open Information Extraction Method Using Cascaded Finite-State Transducer," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2016, pp. 325-337.

[26] Y. Xu, M.-Y. Kim, K. Quinn, R. Goebel, and D. Barbosa, "Open Information Extraction with Tree Kernels," in *HLT-NAACL*, 2013, pp. 868-877.

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

[28] V. Punyakanok, D. Roth, and W.-t. Yih, "The importance of syntactic parsing and inference in semantic role labeling," *Computational Linguistics,* vol. 34, pp. 257-287, 2008.

[29] R. Johansson and P. Nugues, "The effect of syntactic representation on semantic role labeling," in *Proceedings*

وحیده رشادت دوره کارشناسی و ارشد خود را از دانشگاه تبریز در رشته مهندسی کامپیوتر دریافت کرده است. زمینه علاقه‌مندی وی پردازش زبان طبیعی به‌ویژه بازیابی اطلاعات، یادگیری هستان‌نگار و استخراج اطلاعات است.

نشانی رایانامهٔ ایشان عبارت است از:

com.v.reshadat@gmail.com

مریم حورعلی دوره کارشناسی خود را از دانشگاه تهران و کارشناسی ارشد و دکترای خود را از دانشگاه تربیت مدرس در سال ۹۰ دریافت کرده است. ایشان هم‌اکنون استادیار دانشگاه صنعتی مالک اشتر بوده و زمینه علاقه‌مندی وی پردازش زبان طبیعی، خلاصه‌سازی متن، یادگیری هستان‌نگار و تحلیل متن است.

نشانی رایانامهٔ ایشان عبارت است از:

mhourali@mut.ac.ir

هشام فیلی مدرک کارشناسی، کارشناسی ارشد و دکترای خود را از دانشگاه صنعتی شریف به‌ترتیب در سال‌های ۷۶، ۷۸ و ۸۵ دریافت کرده است. ایشان هم‌اکنون دانشیار دانشگاه تهران بوده و زمینه علاقه‌مندی وی هوش مصنوعی، پردازش زبان طبیعی، و تحلیل متن است. وی دارای کتاب‌ها و مقالات بسیاری در نشریات معتبر بین‌المللی است.

نشانی رایانامه ایشان عبارت است از:

hfaili@ut.ac.ir