ارائهٔ روشی برای استخراج اطلاعات ساختاریافته محدود به دامنه از صفحات وب فارسی

حجت امامی

گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه بناب، بناب، ایران

چکیده

استخراج اطلاعات ساختاریافته از متون وب یکی از وظایف اصلی در حوزه وب‌کاوی، پردازش زبان طبیعی و استخراج اطلاعات است. در سال‌های اخیر، روش‌های مختلفی برای استخراج اطلاعات ساختاریافته از متون انگلیسی وب ارائه شده است. اغلب روش‌های موجود برای استخراج اطلاعات در مورد انواع موجودیت‌ها، به یک آنتولوژی از پیش تعریف‌شده نیاز دارند که شامل دانش کامل در مورد موجودیت‌ها و خصلت‌های آن‌ها است. مشکل اصلی این روش‌ها عدم توانایی آن‌ها در استخراج اطلاعات موجودیت‌هایی است که مشخصات آن‌ها از قبل در آنتولوژی تعریف نشده‌اند. در این پژوهش، روش جدیدی برای استخراج خودکار اطلاعات ساختاریافته محدود به دامنه از متون فارسی صفحات وب ارائه شده است که نیازی به دانش پیش‌زمینه در مورد موجودیت‌ها و خصلت‌های آن‌ها ندارد. روش پیشنهادی شامل سه مؤلفه پیش‌پردازش، تحلیل معنایی و نگاشت قاب است. تمرکز اصلی روش پیشنهادی به افزودن اطلاعات معنایی به گزاره‌های مسندآرگومان و استخراج اطلاعات معنادار و محدود به دامنه از گزاره‌ها معطوف شده است. اطلاعات استخراج‌شده در این روش، هم ساختاریافته بوده و هم به مدخل‌های آنتولوژی عمومی DBPedia نگاشت شده‌اند، به‌نحوی‌که پردازش آن‌ها به‌وسیلهٔ ماشین به سهولت انجام می‌شود. برای ارزیابی روش پیشنهادی، یک مجموعه‌دادهٔ کوچک در زبان فارسی ایجاد شده است و روش پیشنهادی و سایر روش‌ها بر روی این مجموعه‌داده مورد ارزیابی قرار گرفته‌اند. نتایج آزمایش‌ها برتری روش پیشنهادی را در مقایسه با سایر روش‌ها برحسب برخی از معیارهای کارایی نشان می‌دهد.

واژگان کلیدی: وب‌کاوی، استخراج اطلاعات، پردازش زبان طبیعی، آنتولوژی، اطلاعات ساختاریافته محدود به دامنه

Presenting a method for extracting structured domain-dependent information from Farsi Web pages

Hojjat Emami

Department of Computer Engineering, Faculty of Engineering, University of Bonab, Bonab, Iran

Abstract

Extracting structured information about entities from web texts is an important task in web mining, natural language processing, and information extraction. Information extraction is useful in many applications including search engines, question-answering systems, recommender systems, machine translation, and etc. An information extraction system aims to identify the entities from the text and extract their related information to form a profile of the target entity.

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

• تاریخ ارسال مقاله: ۱۳۹۸/۱۰/۵ • تاریخ پذیرش: ۱۴۰۰/۳/۳۰ • تاریخ انتشار: ۱۴۰۱/۷/۷ • نوع مطالعه: پژوهشی

In recent years, several methods have been proposed for extracting structured information from web text. The majority of existing methods for extracting entity-centric information require a predefined ontology. The ontology includes the complete knowledge of the entities and their attributes. The main challenge of these methods is their inability to extract information about entities that are not already defined in the ontology. Besides, the existing methods have ignored semantic information extraction and have not linked the extracted information to the general ontology entries. This highlights that introducing new methods for semantic information extraction is an open problem and there is room for more efforts in this field.

As an element of research, we proposed a new method for the automatic extraction of semantically structured information from Farsi web text. The proposed method does not require background knowledge about the entities and their properties. The proposed method consists of three main phases including pre-processing, semantic analysis and frame extraction. To fulfill these phases, we use a combination of language resources, text processing tools, and distant ontologies. The main focuses of the proposed method are to enrich the predicate-argument frames with the semantic information extracted from distant ontologies, extract the entity-related information from predicate-argument frames, and link the extracted information with their corresponding sense in DBPedia ontology. The issue facilitates the processing of Farsi texts by computers.

To evaluate the proposed method, we created a small Farsi dataset containing 100 complete sentences. Then, the proposed method is compared with three information extraction methods on this dataset. The results of experiments show the superiority of the proposed method compared to counterpart methods in terms of precision and F_1 measures.

Keywords: Web mining, information extraction, natural language processing, ontology, structured-semantic information

۱- مقدمه

با افزایش نفوذ اینترنت، توسعه نسخه‌های مختلف وب و بروز رسانه‌های اجتماعی مختلف، هر روز به تعداد کاربران اینترنت افزوده می‌شود [1]. کاربران اینترنت نه‌تنها مصرف‌کننده داده‌ها، بلکه به‌عنوان تولیدکننده داده‌ها بر روی وب، ویکی‌ها، وبلاگ‌ها و سایر رسانه‌های اجتماعی هستند. این موضوع موجب شده تا حجم زیادی از داده‌ها بر روی وب و رسانه‌های اجتماعی مختلف تولید شوند که اغلب به‌صورت غیرساختاریافته هستند [2]. این داده‌ها شامل اطلاعات ارزشمندی در مورد موجودیت‌های مختلف از قبیل اشخاص، مکان‌ها، سازمان‌ها و غیره هستند.

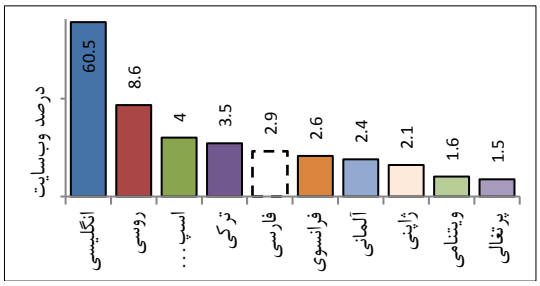
یکی از روش‌های مطلوب کاربران برای دسترسی به اطلاعات موردنظر، استفاده از موتورهای جستجو است. این در حالی است که جستجوی اطلاعات در بین حجم زیادی از متون غیرساختاریافته وب کار بسیار دشواری است. وب‌معنایی به‌عنوان یکی از موضوعات مطرح، در نظر دارد تا از طریق تبدیل داده‌های غیرساختاریافته بر روی وب به دانش ساختاریافته، به هم مرتبط و معنایی، جستجو و دسترسی به اطلاعات را تسهیل کند. یکی از راه‌کارهای مناسب برای تحقق ایده وب‌معنایی استفاده از روش‌های استخراج اطلاعات است. بیش‌تر روش‌های استخراج اطلاعات از یک آنتولوژی[1] از پیش تعریف‌شده برای هدایت فرآیند استخراج اطلاعات استفاده می‌کنند. در این آنتولوژی، موجودیت‌ها و نیز مشخصات آن‌ها به‌صورت دقیق تعریف شده‌اند. این نوع از روش‌های استخراج اطلاعات، به‌عنوان روش‌های ایستا شناخته می‌شوند –[2] [5]. مشکل اصلی روش‌های ایستا نیاز به یک آنتولوژی است که شامل دانش کامل در مورد موجودیت‌ها و خصلت‌های[2] آن‌ها است. این روش‌ها فقط می‌توانند اطلاعات موجودیت‌هایی را استخراج کنند که از قبل در آنتولوژی تعریف شده‌اند و توانایی استخراج اطلاعات جدید و ناشناخته را ندارند. در مقابل روش‌های ایستا، تعداد بسیار محدودی از روش‌ها قرار دارند که می‌توانند اطلاعات محدود به دامنه در مورد موجودیت‌ها را بدون نیاز به آنتولوژی دامنه استخراج کنند. این روش‌ها که به‌عنوان روش‌های استخراج اطلاعات پویا شناخته می‌شوند، توانایی استخراج اطلاعات در مورد انواع مختلف موجودیت‌ها را دارند [8]–[6]. مزیت این روش‌ها عدم نیاز به دانش پیش‌زمینه است، ولی کارایی پایینی دارند و تلاش زیادی برای بهبود کارایی آن‌ها نیاز است. همچنین خروجی این سامانه‌ها دانش سطحی است که به آنتولوژی‌های عمومی که شامل اطلاعات معنایی هستند متصل نشده‌اند.

بیش‌تر روش‌های موجود، برای استخراج اطلاعات از متون صفحات وب انگلیسی ارائه شده‌اند. این در حالی است که در سال‌های اخیر، تعداد کاربران به زبان‌های مختلف به نحو چشم‌گیری افزایش یافته است. این موضوع، موجب افزایش وب‌سایت‌ها به زبان‌های مختلف شده است. در شکل (۱)، زبان‌های مطرح با بیش از ۱/۵٪ سهم تعداد

[1] Ontology

[2] Attributes

وب‌سایت‌ها نشان داده شده است. نمودار شکل (۱) براساس ده میلیون وب‌سایت رسم شده است.



(شکل-۱): درصد وب‌سایت‌ها به زبان‌های مختلف[1]

(Figure-1): Percentages of websites in different languages

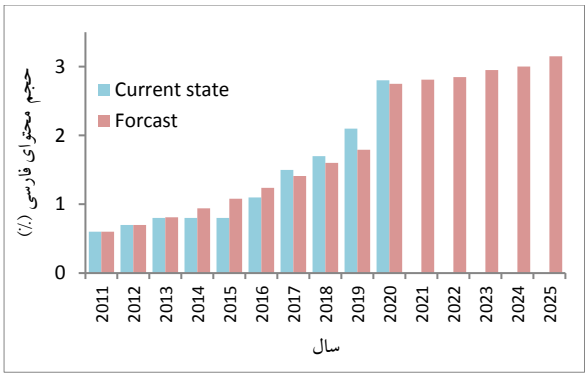
یکی از زبان‌هایی که تلاش اندکی برای استخراج اطلاعات در آن صورت گرفته، زبان فارسی است. در سال‌های اخیر، حجم اطلاعات به زبان فارسی در وب و رسانه‌های اجتماعی مختلف به نحو چشم‌گیری افزایش یافته است و این رشد برای سال‌های آینده نیز پیش‌بینی می‌شود. شکل (۲) رشد حجم اطلاعات را به زبان فارسی در سال‌های اخیر و نیز سال‌های آینده نشان می‌دهد. یکی از دلایل رشد محتوا به زبان فارسی، افزایش ضریب نفوذ اینترنت و گسترش رسانه‌های اجتماعی در کشورهایی است که به زبان فارسی و گونه‌های مختلف آن صحبت می‌کنند. شکل (۳) نقشه پراکندگی فارسی‌زبانان را نشان می‌دهد.

به‌دلیل ماهیت خاص زبان فارسی همانند ساختار زبان و نبود ابزارها و منابع پردازش زبان، پردازش متون فارسی با چالش‌های زیادی مواجه است [9]. این چالش‌ها بر ضرورت توسعه روش‌های استخراج اطلاعات از متون فارسی تأکید می‌کند.

در این پژوهش، روشی معنایی برای استخراج اطلاعات ساختاریافته در مورد موجودیت‌ها از صفحات وب فارسی ارائه شده است. روش پیشنهادی شامل سه مرحله اصلی است که عبارت‌اند از: پیش‌پردازش، تحلیل معنایی و نگاشت قاب. در مرحله پیش‌پردازش صفحات وب پردازش شده و به متن قابل‌پردازش تبدیل می‌شوند. در مرحله تحلیل معنایی، گزاره‌های مسند-آرگومان[2] استخراج شده و با اطلاعات معنایی استخراج‌شده از آنتولوژی عمومی BabelNet غنی‌سازی می‌شوند. در مرحله نگاشت قاب، گزاره‌های مسند-آرگومان به اطلاعات محدود به دامنه نگاشت می‌شوند. خروجی روش پیشنهادی، اطلاعات ساختاریافته معنایی و محدود به دامنه است که تا حد امکان به مدخل‌های آنتولوژی BabelNet و DBPeida متصل شده‌اند.

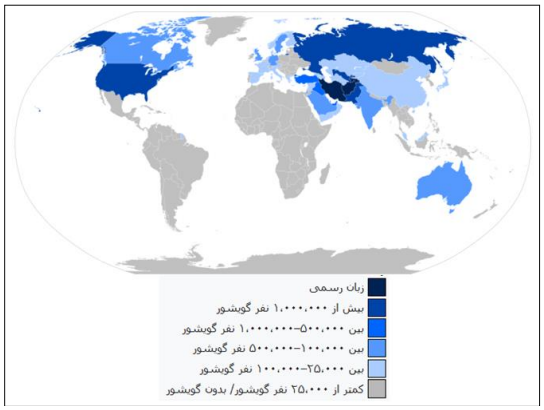
به‌اختصار، نوآوری‌های پژوهش حاضر را می‌توان در موارد زیر خلاصه کرد:

- **ارائه روشی برای استخراج اطلاعات ساختاریافته محدود به دامنه:** در این پژوهش، روشی برای ترکیب ابزارها و منابع مختلف پردازش زبان برای استخراج اطلاعات ساختاریافته ارائه شده است.



(شکل-۲): رشد محتوای فارسی بر روی وب و رسانه‌های اجتماعی[3]

(Figure-2): The growth of Farsi content on the web and social media



(شکل-۳): نقشه پراکندگی فارسی‌زبانان

(Figure-3): Persian language distribution map[4]

در روش پیشنهادی، اطلاعات مربوط به هر موجودیت در قالب یک قاب ساختاریافته بیان شده که

[1] "Web technology surveys (w3techs)", [https://w3techs.com/technologies/overview/content_language], تاریخ آخرین دسترسی: ۲۰ آذر ۱۳۹۹

[2] Predicate-argument propositions

[3] داده‌های مورد استفاده برای رسم وضعیت فعلی محتوای فارسی بر روی اینترنت، از وب‌سایت [www.W3Techs.com] استخراج شده است. تاریخ آخرین دسترسی: ۲۰ آذر ۱۳۹۹

[4] https://fa.wikipedia.org/wiki/زبان فارسی

مدخل‌های این قاب معنادار بوده و خصلت‌های آن موجودیت را نشان می‌دهد.

- **تخصیص معنا به متن سطحی:** در این روش، اطلاعات استخراج‌شده تا حد امکان به مدخل‌های آنتولوژی‌های عمومی BabelNet و DBPeida متصل شده‌اند تا هم خوانایی مناسبی داشته باشند و هم پردازش آن‌ها برای رایانه تسهیل شوند. روش پیشنهادی در تحقق ایدهٔ وب‌معنایی تا حد زیادی مفید است.
- **ارزیابی روش پیشنهادی:** یکی از چالش‌های اساسی در ارزیابی روش پیشنهادی، عدم وجود مجموعه‌داده استاندارد در زبان فارسی است؛ به همین منظور، یک مجموعه‌داده کوچک ایجاد و روش پیشنهادی بر اساس معیارهای مختلف بر روی آن ارزیابی شده است. نتایج آزمایش‌ها، برتری روش پیشنهادی را در مقایسه با سایر روش‌ها نشان می‌دهد.

پس از این مقدمه کوتاه، ساختار مقاله بدین صورت سازمان‌دهی شده است: در بخش ۲، کارهای مرتبط در زمینهٔ استخراج اطلاعات از متون غیرساختاریافته بررسی و در بخش ۳، به جزئیات روش پیشنهادی پرداخته شده است. ارزیابی روش پیشنهادی به‌همراه نتایج آزمایش‌ها و بحث بر روی نتایج در بخش ۴ آورده شده و نتیجه‌گیری مقاله و نیز ارائه پیشنهادهایی برای انجام پژوهش‌های بیشتر در بخش ۵ آورده شده است.

۲- کارهای مرتبط

در سال‌های اخیر، روش‌های متعددی برای استخراج اطلاعات ساختاریافته از متون زبان طبیعی غیرساختاریافته ارائه شده است. روش‌های اولیه اغلب مبتنی بر تطابق الگو هستند. این روش‌ها از الگوهای نحوی که توسط افراد خبره طراحی می‌شوند، برای جستجوی متن و تطابق رشته‌ها استفاده می‌کنند. به‌عنوان‌مثال، لی[1] و همکارانش [3] از سلسله‌مراتبی از الگوهای نحوی و زبانی برای استخراج اطلاعات در مورد افراد، سازمان‌ها و مکان‌های مختلف استفاده کردند. در مرجع [10]، برای استخراج خصلت‌های اشخاص، از روش تطابق الگو و نیز کلیدواژه‌های مرتبط با خصلت‌های موردنظر استفاده کردند. در این روش، ابتدا با یک سامانه تشخیص موجودیت نامدار، عبارات نامبری[2] از متن استخراج می‌شوند، اگر کلیدواژه‌ای که مبین خصلت موردنظر است در جمله وجود داشته باشد، عبارت نامبری به‌عنوان مقدار آن خصلت در نظر گرفته می‌شود. به‌عنوان‌مثال، در جمله «Alice was born in Tokyo.» کلمه "Tokyo" به‌عنوان «محل تولد» شخص "Alice" در نظر گرفته می‌شود، زیرا با کلیدواژه "born" به یکدیگر مرتبط شده‌اند. در مرجع [11]، روشی برای استخراج اطلاعات از صفحات وب با استفاده از قالب‌های محدود به دامنه ارائه شده است. در این روش، برای استخراج اطلاعات از الگوهای نحوی استفاده شده و نتایج در قالب فایل‌های XML ذخیره شده است. در مرجع [5] روشی برای استخراج روابط محدود به دامنه ارائه شده که بر اساس روش‌های مبتنی بر دانش و یادگیری ماشینی از راه دور[3] است. در این روش، از آنتولوژی‌های DBPedia و Freebase برای استخراج روابط بین موجودیت‌ها استفاده شده است. مشکل اصلی این روش نیاز آن‌ها به آنتولوژی دامنه برای هدایت فرآیند استخراج اطلاعات است. آنتولوژی شامل مجموعه‌ای از مفاهیم و روابط بین آن‌ها در یک حوزه خاص است.

در مقابل روش‌های مبتنی بر آنتولوژی، روش‌هایی وجود دارند که برای استخراج اطلاعات ساختاریافته از متن بدون نیاز به آنتولوژی دامنه ارائه شده‌اند [15]–[12]. چامبرز[4] و جورافسکی[5] [15]، روشی برای استخراج اطلاعات ساختاریافته از متون خام بدون نیاز به آنتولوژی ارائه دادند. آن‌ها وظیفه استخراج اطلاعات موجودیت‌ها را به‌عنوان مسأله استخراج مجموعه‌ای از <خصلت، مقدار خصلت> فرموله‌بندی کرده‌اند. هدف آن‌ها یافتن عناوین خصلت‌ها و مقادیر مناسب برای آن خصلت‌ها است. آن‌ها برای یادگیری خصلت‌ها، از گروه‌بندی افعال کنشی به هم مرتبط استفاده کردند. همچنین برای استخراج مقادیر مناسب برای یک خصلت، نقش‌های دستوری افعال موجود در گروه مربوط به آن خصلت را گروه‌بندی کرده و به این طریق مقادیر مناسب را برای آن پیدا کردند. چالش اصلی این روش آن است که مقادیر خصلت‌ها به‌صورت متن سطحی هستند. منظور از متن سطحی، متنی است که به‌صورت دنباله‌ای از رشته‌ها است که به مدخل‌های یک آنتولوژی متصل نشده‌اند.

مرجع [12] یک راهکار بدون نظارت برای استخراج اطلاعات ساختاریافته و نمایش آن‌ها در قالب یک جدول ارائه می‌کند. این روش، یک مدل استخراج رابطه است که به یک موجودیت خاص محدود نیست و فقط دنباله‌ای از روابط را از متن کشف می‌کند. در این روش، استخراج

[1] Li
[2] Mention
[3] Distant machine-learning
[4] Chambers
[5] Jurafsky

روابط به‌صورت یک مسأله خوشه‌بندی فرموله‌بندی شده است. ابتدا اسناد مختلف که در مورد یک رخداد یکسان هست خوشه‌بندی شده است؛ سپس عبارات پرتکرار که عبارت اسمی یکسانی را به هم متصل می‌کنند، به‌عنوان روابط بین عبارات اسمی استخراج می‌شوند. محدودیت این روش این است که اطلاعات استخراج‌شده روابط آزاد بوده و محدود به مفاهیم دامنه خاصی نیستند.

بانکو[1] و همکارانش [16] سیستم TextRunner را معرفی کردند که قادر است روابط آزاد را بدون استفاده از دانش پیش‌زمینه در مورد موجودیت‌ها استخراج کند. سیستم TextRunner فقط می‌تواند مجموعه‌ای از روابط را استخراج کند که به دامنه خاصی محدود نیستند. کاری که در مراجع [7] و [8] انجام شده تلاشی در جهت محدودکردن روابط آزاد به دانش محدود به دامنه است. در این روش‌ها، برای نگاشت هر رابطه آزاد به یک رابطهٔ محدود به دامنه، از مجموعه‌ی از قواعد نگاشت روابط و نیز فهرستی از کلمات کلیدی مختص آن دامنه استفاده شده است. محدودیت این روش‌ها این است که روابط استخراج‌شده فقط رشته‌های متنی هستند که فاقد معنا بوده و رفع ابهام نشده‌اند.

برخی دیگر از پژوهش‌گران، از منابع پردازش زبان استفاده و تلاش کرده‌اند تا اطلاعات ساختاریافته را به قالب FrameNet[2] استخراج کنند [19]–[17]. این روش‌ها بیش‌تر از شیوه‌های دسته‌بندی برای استخراج خصلت‌های موجودیت‌ها استفاده کرده‌اند. برخلاف این‌که این روش‌ها توانایی استخراج اطلاعات ساختاریافته را دارند، ولی این اطلاعات به دانش دامنه محدود نشده و به مدخل‌های آنتولوژی‌های عمومی نگاشت نشده‌اند؛ بنابراین نیاز است تا این اطلاعات بیشتر پردازش شوند تا برای تحقق ایده وب‌معنایی مناسب باشند.

اغلب روش‌ها برای استخراج اطلاعات از متون انگلیسی ارائه شده و تنها تعداد محدودی از این روش‌ها بر روی زبان فارسی تمرکز کرده‌اند. روش‌های ارائه‌شده برای استخراج اطلاعات در زبان فارسی، فقط توانایی استخراج اطلاعات در مورد موجودیت‌هایی را دارند که از قبل مشخصات آن‌ها در آنتولوژی دامنهٔ موردنظر تعریف شده‌اند. اغلب روش‌های ارائه‌شده برای استخراج اطلاعات در زبان فارسی، از الگوهای نحوی و لغوی استفاده می‌کنند [23]–[20]. این روش‌ها که اغلب بر مبنای روش هرست[3] هستند، کار استخراج اطلاعات را در دو مرحله انجاممی‌دهند. در مرحله نخست، با استفاده از الگوهای مختلف که متکی بر کلمات خاص و نیز تجزیه نحوی متن هستند، اطلاعات موردنظر استخراج می‌شوند. در مرحله دوم، به جهت افزایش دقت اطلاعات استخراج‌شده، پردازش‌هایی برای پالایش اطلاعات صورت می‌گیرد. این روش‌ها از چندین چالش اساسی رنج می‌برند: تنوع لغوی و نحوی در آن‌ها در نظر گرفته نشده است؛ اطلاعات استخراج‌شده به مدخل‌های آنتولوژی متصل نشده و متون سطحی هستند؛ و مشکلات خاص زبان همانند هم‌معنایی و چندمعنایی کلمات در نظر گرفته نشده است. برای رفع این مشکلات، در مرجع [2] روشی برای استخراج اطلاعات موجودیت‌ها از متون غیرساختاریافته ارائه شده است. این روش بر تحلیل معنایی متن و غنی‌سازی اطلاعات سطحی با اطلاعات معنایی استخراج‌شده از آنتولوژی BebelNet استوار است. اطلاعات استخراج‌شده به مدخل‌های آنتولوژی عمومی DBPedia متصل شده‌اند و ساختار مشخصی دارند. مشکل اصلی این روش انعطاف‌پذیری پایین آن است؛ زیرا متکی به یک آنتولوژی دامنه است و تنها می‌تواند اطلاعاتی را که از قبل در آنتولوژی دامنهٔ موردنظر تعریف شده‌اند، استخراج کند. روش پیشنهادی در این پژوهش، مشکل وابستگی به آنتولوژی دامنه را حل کرده و توانایی استخراج اطلاعات ساختاریافته و معنادار را دارد.

۳- روش پیشنهادی

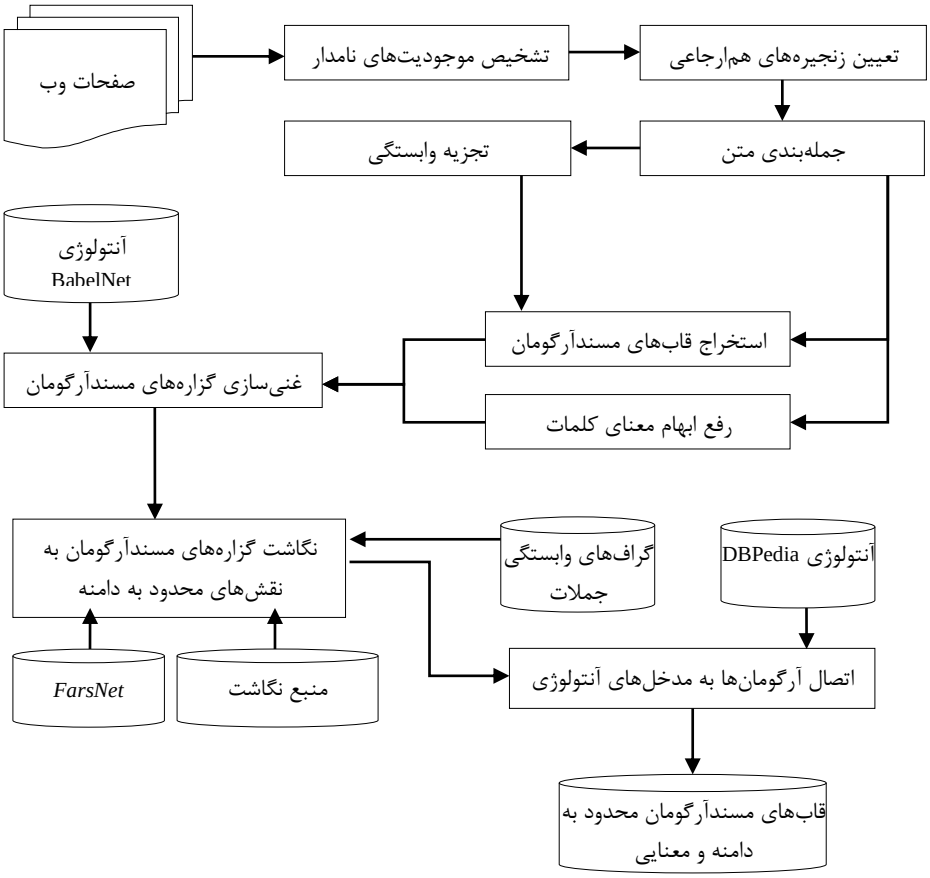
در این پژوهش، مسأله استخراج اطلاعات ساختاریافته معنایی از متون فارسی غیرساختاریافته وب موردبررسی قرار گرفته است. با فرض اینکه $D=\{D_1, D_2, ..., D_n\}$ نشان‌دهنده مجموعه صفحات وب باشد، هدف روش پیشنهادی، استخراج اطلاعات ساختاریافته و معنایی در یک قالب ساختاریافته به‌صورت $T=\{<a_1, v_1, s_1>, <a_2, v_2, s_2>, ..., <a_k, v_k, s_k>\}$ است که در آن a_i بیان‌گر عنوان خصلت، v_i بیان‌گر مقدار خصلت و s_i بیان‌گر شناسه مدخل Babelfy[4] است که مقدار خصلت سطحی v_i را به یک معنای عمقی منحصربه‌فرد در آنتولوژی BabelNet نگاشت می‌کند. شکل (۴) ساختار روش پیشنهادی برای استخراج اطلاعات ساختاریافته از متون وب را نشان می‌دهد.

روش پیشنهادی شامل سه مؤلفه اصلی است:

[1] Banko
[2] https://framenet.icsi.berkeley.edu/fndrupal/
[3] Hearst
[4] http://babelfy.org/

- **پیش‌پردازش:** در این مرحله، صفحات وب با استفاده از ابزارهای پیش‌پردازش به قالب مناسب برای پردازش‌های آینده تبدیل می‌شوند. تمرکز اصلی در این پژوهش، بر روی متون صفحات وب است؛ زیرا اغلب اطلاعات باارزش در مورد موجودیت‌ها در قالب متن بیان می‌شوند. به همین جهت، صفحات وب باید پردازش و متن موجود در آن‌ها به فرمت مناسب تبدیل شوند.
- **تحلیل معنایی متن:** این مؤلفه، قاب‌های مسندآرگومان را از هر جملهٔ موجود در متن پیش‌پردازش‌شده استخراج می‌کند. تحلیل معنایی شامل سه مؤلفه رفع ابهام معنای کلمات، استخراج قاب‌های مسندآرگومان و غنی‌سازی معنایی است.
- **استخراج قاب ساختاریافته:** این مؤلفه، قاب‌های ساختاریافته محدود به دامنه را از قاب‌های مسندآرگومان مستقل از دامنه استخراج می‌کند. این مؤلفه شامل دو مرحله نگاشت قاب و اتصال به آنتولوژی DBPedia است.

در ادامه، این مؤلفه‌ها به‌تفصیل موردبررسی قرار گرفته‌اند.



(شکل-۴): روش پیشنهادی برای استخراج اطلاعات ساختاریافته معنایی از متون فارسی
(Figure-4): The proposed method for extracting semantic structured information from Farsi text

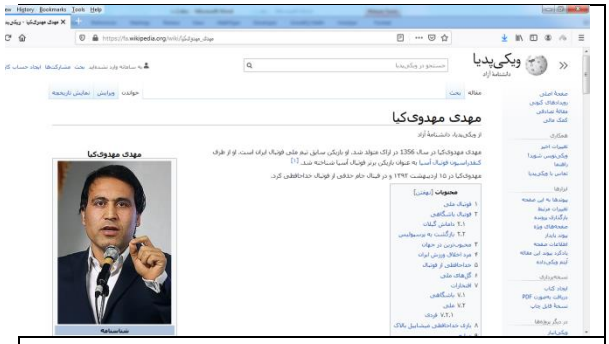
۱-۳- پیش‌پردازش

به‌منظور تسهیل در پردازش صفحات وب، نیاز است تا چهار مرحله زیر به‌ترتیب انجام شوند: (۱) حذف تگ‌های HTML و استخراج متن خام از صفحات وب، (۲) تشخیص موجودیت‌های نامدار، (۳) تحلیل هم‌ارجاعی و (۴) جمله‌بندی متن. در مرحله نخست، با استفاده از ابزار Jsoup[1] تگ‌های HTML از صفحات وب حذف شده و به اسناد متنی تبدیل می‌شوند؛ سپس، با استفاده از ابزار Polyglot-NER [24] انواع مختلف موجودیت‌ها شامل شخص، مکان و سازمان از هر سند استخراج می‌شوند؛ سپس سند متنی به سیستم تحلیل هم‌ارجاعی داده می‌شود تا زنجیره عبارات نامبری هم‌مرجع مشخص شوند. برای انجام این کار از یک ابزار تحلیل هم‌ارجاعی مبتنی بر قانون که در مرجع [25] ارائه شده، استفاده شده است. پس از تعیین زنجیره‌های هم ارجاعی، نامبری‌های موجود در هر زنجیره با نامبری نماینده آن زنجیره جایگزین می‌شوند. در مرحله جمله‌بندی، با ابزار جمله‌بند هضم[2]، متن به مجموعه‌ای از جملات تقسیم می‌شود. ابزار تشخیص‌دهندهٔ جملات با استفاده از علامت‌های پایانی جمله و به‌کارگیری برخی دستورهای گرامری زبان فارسی

[1] Jsoup: Java HTML parser, [http://jsoup.org/]

[2] http://www.sobhe.ir/hazm/

و درنظرگرفتن حروف ربط و برخی لغات آغازکنندهٔ جملات مرکب، مرز جمله‌ها را تعیین می‌کند. در مرحله جمله‌بندی، فقط جملاتی در نظر گرفته شده‌اند که دارای یک عبارت فعلی و دست‌کم یک عبارت اسمی به‌عنوان موجودیت نامدار هستند. شکل (۵) نتیجه مرحله پیش‌پردازش را برای صفحه وب مربوط به موجودیت «مهدی مهدوی‌کیا» نشان می‌دهد.



جمله ۱: مهدی مهدوی‌کیا در سال ۱۳۵۶ در اراک متولد شد.

جمله ۲: مهدی مهدوی‌کیا بازیکن سابق تیم ملی فوتبال ایران است.

جمله ۳: مهدی مهدوی‌کیا از طرف کنفدراسیون فوتبال آسیا به عنوان بازیکن برتر فوتبال آسیا شناخته شد.

جمله ۴: مهدی مهدوی‌کیا در ۱۵ اردیبهشت ۱۳۹۲ و در فینال جام حذفی از فوتبال خداحافظی کرد.

(شکل-۵): نتیجه مرحله پیش‌پردازش بر روی یک صفحه وب نمونه

(Figure-5): Result of the pre-processing phase on a sample web page

۲-۳- تحلیل معنایی متن

این مؤلفه، متن پیش‌پردازش‌شده را به‌عنوان ورودی گرفته، سپس برای هر جمله از متن، قاب‌های مسند آرگومان را تولید کرده و آن‌ها را با اطلاعات معنایی غنی‌سازی می‌کند. خروجی مؤلفه تحلیل معنایی، قاب‌های مسندآرگومان غنی‌شده با اطلاعات معنایی است. تحلیل معنایی شامل سه مرحله اصلی است که عبارت‌اند از: رفع ابهام معنای کلمه[1]، استخراج قاب‌های مسندآرگومان و غنی‌سازی معنایی[2].

۱-۳-۲- رفع ابهام معنای کلمه

در این مرحله، کلمات و نامبری[3] موجودیت‌ها در متن به مدخل‌های متناظر آن‌ها در آنتولوژی BabelNet نگاشت می‌شوند. برای انجام این کار، با استفاده از ابزار Babelfy [26]، کلمات سطحی متن رفع ابهام شده و به معنی[4] واقعی آن‌ها در آنتولوژی BabelNet نگاشت می‌شوند. اگر معنای یک موجودیت در BabelNet موجود نباشد، روش پیشنهادی امکان رفع ابهام آن را نخواهد داشت، و این یکی از ایرادهای روش پیشنهادی است. مثال ۱ نتیجه رفع ابهام معنایی برای جمله (۱) آورده شده در شکل (۵) را نشان می‌دهد. bn:*in* نشان‌دهنده کد معنی *i* ام در BabelNet برای کلمه موردنظر است.

مثال ۱:

مهدی مهدوی‌کیا	←	*bn:02088605n*
در	←	-
سال	←	-
۱۳۵۶	←	
در	←	-
اراک	←	*bn:00257292n*
متولد شد.	←	-

۲-۳-۲- استخراج قاب‌های مسندآرگومان

این مؤلفه، متن ورودی را به مجموعه‌ای از قاب‌های مسند آرگومان تبدیل می‌کند. فرض کنید $P=\{v, E_1, E_2, ..., E_k\}$ نشانگر قاب مسندآرگومان برای یک جمله s باشد. در قاب P، متغیر v بیان‌گر فعل جمله و $E_i=(s_i, a_i)$ عنصر i در قاب است که در آن s_i بیان‌گر نقش معنایی برای آرگومان a_i است. جدول (۲) فهرست نقش‌های معنایی را که در این پژوهش در نظر گرفته شده‌اند، نشان می‌دهد.

(جدول-۲): فهرست نقش‌های معنایی

(Table-2): List of semantic roles

کنشگر (Agent)	مبدأ (Source or Origin)
آزمایشگر (Expriencer)	زمان (Time)
موضوع (Theme)	بهره‌بردار (Beneficiary)
کنش‌پذیر (Patient)	روش (Manner)
ابزار (Instrument)	هدف (Purpose)
مقصد (Goal)	محتوا (Content)

برای استخراج گزاره‌های مسندآرگومان از یک روش مبتنی بر قاعده [27] استفاده شده است. برحسب تعداد فعل‌های موجود در جمله، ممکن است یک یا چندین گزاره مسندآرگومان در یک جمله وجود داشته باشد. مجموعه نقش‌های معنایی در گزاره P به معنای فعل جمله وابسته بوده و مستقل از دامنه هستند [28]، همانند

[1] Word sense disambiguation
[2] Semantic enrichment
[3] Mention
[4] Sense

نقش‌های کنش‌گر، کنش‌پذیر و موضوع. قاب مسندآرگومان برای جمله شماره (۱) در شکل (۵) به‌صورت زیر است.

مثال (۲):

P= {زمان> ,< مهدی مهدوی‌کیا ,کنشگر> ,متولد شدن
,<اراک ,مکان> ,<۱۳۵۶}

۳-۳-۲- غنی‌سازی معنایی

این مؤلفه، قاب‌های مسندآرگومان را به‌عنوان ورودی دریافت کرده و آن‌ها را با اطلاعات معنایی که در مرحله رفع ابهام معنای کلمه تولید شده‌اند، غنی‌سازی می‌کند. برای غنی‌سازی قاب‌های مسندآرگومان، هرکدام از آرگومان‌ها به معنای معادل آن در آنتولوژی BabelNet مرتبط می‌شوند. در هنگام غنی‌سازی، اگر عبارات موجود در یک آرگومان با چندین مدخل در آنتولوژی BabelNet متناظر باشد، در این صورت عبارات موجود در آرگومان به چندین مدخل آنتولوژی مرتبط می‌شوند. اگر عبارات موجود در یک آرگومان به یک موجودیت و یا مفهوم یکسان اشاره کند و یا اگر این عبارات هم مرجع باشند، در این صورت به‌جای چندین معنی برای عبارات، تنها از یک معنی که به‌صورت تصادفی انتخاب می‌شود، استفاده می‌شود. به‌عنوان‌مثال، در جمله «حسن روحانی، رئیس‌جمهور ایران در سرخه زاده شد.» می‌توان به‌جای استفاده از دو معنی bn:14918092n برای عبارت حسن روحانی و bn:01650357n برای رئیس‌جمهور ایران، از معنی bn:14918092n استفاده کرد. عناصر باقی‌مانده که هیچ معادلی در مجموعه معانی رفع ابهام شده ندارند، بدون هیچ‌گونه تغییری باقی می‌مانند.

۳-۳- نگاشت قاب

در این مرحله، قاب‌های مسندآرگومان مستقل از دامنه به قاب‌های مختص دامنه تبدیل می‌شوند. در زبان انگلیسی، برای انجام این کار می‌توان از منبع نگاشت SemLink [29] استفاده کرد که منبعی برای نگاشت نقش‌های مستقل دامنه از فرمت VerbNet به نقش‌های محدود به دامنه به فرمت FrameNet است. استفاده از این منبع در زبان فارسی مقدور نیست. در این پژوهش، با الهام از روش ارائه‌شده در مرجع [29] و نیز با استفاده از قاب‌هایی که در FarsNet[1] وجود دارند، منبع نگاشت کوچکی برای تبدیل نقش‌های مستقل از دامنه به نقش‌های محدود به دامنه ارائه شده است. منبع ایجادشده شامل مجموعه‌ای از قاب‌ها در قالب XML است. به‌دلیل محدودیت در برچسب‌گذاری قاب‌ها، منبع ایجادشده فقط شامل یکصد قاب نگاشت است. شکل (۶) یک نمونه از قاب نگاشت نقش را برای مصدر «متولدشدن» نشان می‌دهد.

```
< frame>
 <infinitive> متولد شدن </infinitive>
 <meaning> بیرون آمدن نوزاد از رحم مادر</meaning>
 <category> NP; NP; NP </category>
 <syntacticRole>فاعل; قید; قید </syntacticRole>
 <argmap>
    <P-role ="کنشگر",  Fn-role="نام"/>
    < P-role ="زمان",    Fn-role ="زمان تولد"/>
    < P-role ="مکان",    Fn-role ="محل تولد"/>
 </argmap>
<id_sense_farsnet>22371</id_sense_farsnet>
</frame>
```

(شکل-۶): نمونه‌ای از قاب نگاشت نقش برای مصدر «متولدشدن»

(Figure-6): An example of role mapping frame for verb "being born"

برای نگاشت نقش قاب‌های مسندآرگومان به نقش‌های محدود به دامنه نیاز است دو مرحله صورت گیرد: تعیین قاب[2] و نگاشت نقش[3]. مرحله تعیین قاب، مسئول تعیین قاب مناسب برای یک فعل v است. مؤلفه نگاشتِ نقش، مسئول نگاشت آرگومان‌های مسندآرگومان به نقش‌های معنایی محدود به دامنه است.

۳-۳-۱- تعیین قاب

با فرض اینکه $Fn=\{F_1, F_2, ..., F_k\}$ قاب‌های نگاشت باشد، برای پیداکردن قاب نگاشت مناسب از روش تطابق رشته استفاده می‌شود. به این منظور، فعل قاب مسند آرگومان P با افعال قاب‌های Fn مقایسه شده و درصورتی که تطابقی پیدا شود، قاب $F_i \in Fn$ به‌عنوان قاب نامزد برای نگاشت نقش‌ها در نظر گرفته می‌شود. برای افعالی که نمی‌توان به‌صورت مستقیم، قاب معادلی در Fn پیدا کرد، از مترادف افعال برای پیداکردن قاب نگاشت مناسب استفاده می‌شود. مترادف افعال نشان‌گرهای خوبی برای پیداکردن قاب متناظر یک فعل هستند؛ زیرا یک فعل بر اساس معنای آن به یک قاب نگاشت می‌شود. پس از تعیین افعال مترادف فعل v، قاب‌های متناظر این افعال از Fn استخراج می‌شوند. از بین قاب‌های استخراج‌شده، قابی انتخاب می‌شود که عناصر آن بیش‌ترین تطابق را با عناصر قاب مسندآرگومان داشته باشند. همچنین می‌توان قاب متناظر را به‌صورت تصادفی نیز انتخاب کرد.

[1] http://farsnet.nlp.sbu.ac.ir/Site3/Modules/Public/Default.jsp

[2] Frame mapping
[3] Role mapping

۲-۳-۳- نگاشت نقش

مؤلفه نگاشت نقش، قاب مسندآرگومان P و بردار قاب نامزد F_i را به‌عنوان ورودی دریافت می‌کند و نقش‌های عمومی در قاب P را به نقش‌های مختص دامنه نگاشت می‌کند. نگاشت نقش برای قاب مسندآرگومان P، از طریق مقایسه مدخل‌های قاب P و عناصر قاب F_i انجام می‌شود. برای انجام این کار، از دو نوع محدودیت استفاده می‌شود: محدودیت نقش دستوری[1] (GF) و نوع عبارت[2] (PT). نقش‌های دستوری و نوع عبارت جزو ویژگی‌های نحوی[3] هستند. نقش دستوری نشان‌دهنده نقش نحوی (مانند فاعل، فعل، مفعول) یک عبارت است. نوع عبارت، برچسب جز کلام[4] (مانند عبارت اسمی، فعلی) یک عبارت را نشان می‌دهد. این محدودیت‌ها دقت نگاشت نقش را افزایش می‌دهد. برای تعیین نقش دستوری و نوع عبارت از پارسر هضم استفاده شده است. اگر مؤلفه استخراج قاب‌های مسندآرگومان برخی از نقش‌ها را به‌صورت نادرست استخراج کرده باشد، استفاده از محدودیت‌های نقش دستوری و نوع عبارت خطای نگاشت نقش‌ها کاهش می‌یابد. مقدار آرگومان a_i برای نقش s_i در قاب P، با مقدار آرگومان m_j نقش e_j در قاب F_i مقایسه می‌شود. اگر دو آرگومان a_i و m_j بر اساس محدودیت‌های نقش دستوری و نوع عبارت یکسان باشند، در این صورت نقش e_j به آرگومان a_i اختصاص داده می‌شود. در برخی موارد، فعل موجود در گزاره‌های مسندآرگومان P متناظری در مجموعه قاب‌های Fn ندارد، مترادف آن فعل در Fn وجود ندارد و یا قاب P آرگومان‌هایی دارد که در قاب $F_i \in Fn$ متناظری ندارند. در این‌گونه موارد، آرگومان‌های قاب مسندآرگومان بدون نگاشت به فرم اولیه خود رها می‌شوند. الگوریتم (۱) نگاشت عناصر قاب مسندآرگومان را به نقش‌های محدود به دامنه نشان می‌دهد.

نتایج نگاشت نقش برای قاب مسندآرگومان آورده شده در مثال (۲) به‌صورت زیر است.

مثال (۳):

```
<Frame docID="1" sentID="1" verb="متولد شدن" sense="22371"
>
  <argmap>
      <role P-role="کنشگر" Fn-role="نام" value="مهدی مهدوی‌کیا"
/>
      <role P-roe="زمان" Fn-role="تاریخ تولد" value="۱۳۵۶ " />
      <role P-role="مکان" Fn-role="محل تولد" value="اراک" />
  </argmap>
</Frame >
```

همان‌گونه که در مثال (۳) نشان داده شده است، نقش «کنش‌گر» به نقش محدود به دامنه «نام فرد» و نقش‌های «زمان» و «مکان» به‌ترتیب به نقش‌های «تاریخ تولد» و «محل تولد» نگاشت شده‌اند.

(الگوریتم-۱): الگوریتم نگاشت عناصر قاب مسندآرگومان به نقش‌های محدود به دامنه

(Algorithm-1): The algorithm of mapping predicate-argument frame's elements to domain-dependent roles

Input:
P: Predicate-argument frame
sn: Sense of the verb *v*
F_i: Candidate frame in mapping source

Output:
Frame with domain-dependent roles

Procedure:

for each thematic role $s_i \in P$
  a_i = argument of s_i ;
  for each semantic role $e_j \in F_i$
    m_j = argument of e_j
    if ($GF(a_i) = GF(m_j)$ &
      $PT(a_i) = PT(m_j)$)
        role (a_i)= e_j ;
    else
        role (a_i)= original thematic role s_i
    end if

۴-۳- اتصال به آنتولوژی DBPedia

در این مرحله، آرگومان‌های موجود در قاب‌های مسندآرگومان به مدخل‌های آنتولوژی DBPedia متصل می‌شوند. برای انجام این کار، URI متناظرِ هر آرگومان در DBPedia با استفاده از معنی آن در BabelNet مشخص می‌شود، درنهایت، آرگومان موردنظر به مدخل متناظر آن در DBPedia متصل می‌شود. مثال زیر نتایج اتصال آرگومان‌ها به مدخل‌های متناظر آن‌ها در آنتولوژی DBPedia را نشان می‌دهد.

مثال (۴):

```
<Frame docID="1" sentID="1" verb="متولد شدن" sense="22371"
>
    <elements>
        <element Fn-role="نام" value="مهدی مهدوی‌کیا "
            babelfyID=" bn:02088605n " DBPediaURI="
            http://dbpedia.org/page/Mehdi_Mahdavikia" />
        <element Fn-role="تاریخ تولد" fnrole="Partner_2"
```

[1] Grammatical function
[2] Phrase type
[3] Syntactic features
[4] Part of speech (POS)

```
        value="۱۳۵۶ "
            babelfyID="- " DBPediaURI="-" />
        <element Fn-role="محل تولد" fnrole="Place"
        value="اراک"
            babelfyID=" bn:00257292n " DBPediaURI="
         http://dbpedia.org/page/Arak,_Iran" />
    </elements >
</Frame >
```

۴- ارزیابی روش پیشنهادی

۴-۱- معیارهای کارایی

کارایی روش پیشنهادی با استفاده از سه معیار دقت[1]، بازخوانی[2] و میانگین همساز دقت و بازخوانی (F_1) [30] ارزیابی شده است. کار ارزیابی در دو بخش صورت گرفته است: ۱) توانایی روش پیشنهادی در استخراج قاب‌ها از متن و ۲) توانایی روش پیشنهادی در استخراج محتوای قاب‌ها که شامل زوج مرتب‌های <عنوان خصلت و مقدار خصلت> است. معیارهای کارایی به‌واسطه مقایسه نتایج تولیدشده به‌وسیلهٔ سیستم و مجموعه استاندارد طلایی محاسبه شده است. به‌صورت ریاضی، معیارهای دقت (P)، بازخوانی (R) و معیار F_1 به‌صورت زیر تعریف می‌شود:

$$P = \frac{|S \cap G|}{|S|} \quad (۴)$$

$$R = \frac{|S \cap G|}{|G|} \quad (۵)$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \quad (۶)$$

در این رابطه، S نشان‌گر مجموعه قاب‌ها و زوج مرتب‌های <خصلت، مقدار خصلت> است که به‌وسیله سیستم استخراج شده است و G مجموعه قاب‌ها و زوج‌مرتب‌های <خصلت، مقدار خصلت> در مجموعه استاندارد طلایی است.

۴-۲- مجموعه‌داده

یکی از چالش‌های اساسی در ارزیابی روش پیشنهادی، عدم وجود مجموعه داده مناسب در زبان فارسی است. به همین منظور، در این پژوهش، یک مجموعه‌داده کوچک برای ارزیابی روش پیشنهادی ایجاد کرده‌ایم. برای ایجاد مجموعه‌داده، یکصد جمله کامل غیرساختاریافته که دست‌کم شامل یک فعل کنشی اصلی هستند، از ویکی‌پدیا[3] و سایر رسانه‌ها انتخاب شده‌اند. برخی از جملات انتخاب‌شده مورد بازبینی قرار گرفته و تغییر داده شده‌اند تا دست‌کم دارای یک قاب باشند. این جملات با ساختارهای نحوی متفاوت بیان شده، شامل موجودیت‌های مختلف بوده و در حوزه‌های مختلفی هستند. برای ایجاد مجموعه استاندارد طلایی، به‌کمک دو خبره انسانی قاب‌های ساختاریافته متناظر جملات مشخص شدند. مجموعه‌داده مورد استفاده به نشانی زیر قابل دسترس است.

https://www.ubonab.ac.ir/uploads/user/310/files/dataset.doc

گفتنی است با توجه به محدودبودن اندازه مجموعه داده نمی‌توان ادعا کرد که ارزیابی روش پیشنهادی به میزان زیادی قابل‌اعتماد است. جدول (۳) مشخصات مجموعه‌داده مورداستفاده را نشان می‌دهد. مجموعه‌داده‌ای که برای ارزیابی مسأله استخراج قاب‌های ساختاریافته در زبان انگلیسی در کارگاه SemEval ارائه شده است، شامل ۱۲۰ جمله است که مقیاس کوچکی دارد [31]. در این مسأله، استخراج تعداد قاب‌ها مطرح نبوده و کیفیت استخراج قاب‌ها و نیز مدخل‌های آن‌ها مطرح است. امّا اندازه مجموعه‌داده موردتوجه نبوده است. ایجاد مجموعه‌داده بزرگ‌تر و منسجم‌تر برای ارزیابی عملکرد روش پیشنهادی جزو کارهای آینده در نظر گرفته شده است.

(جدول-۳): مشخصات مجموعه‌داده آزمون
(Table-3): The characteristics of test dataset

تعداد جمله‌ها	تعداد قاب‌ها	تعداد موجودیت‌های نامدار
۱۰۰	۱۰۰	۲۰۳

۴-۳- نتایج

برای ارزیابی روش پیشنهادی، سه روش دیگر که برای استخراج اطلاعات ساختاریافته از متون انگلیسی پیشنهاد شده‌اند، پیاده‌سازی شده و بر روی مجموعه‌داده فارسی مورد ارزیابی قرار گرفته است. این روش‌ها عبارت‌اند از: UTD-SRL [19] روش AFE [18] و روش چامبرز [15].

روش UTD-SRL [19] از ترکیب دو دسته‌بند برای استخراج قاب و محتویات آن‌ها استفاده می‌کند: ماشین بردار پشتیبان و بیشینه آنتروپی. این روش از دقت بالایی در تشخیص قاب و عناصر قاب‌ها برخوردار است. روش AFE [18] از الگوریتم‌های دسته‌بندی که بر روی گراف وابستگی جملات اعمال می‌شوند برای تشخیص قاب‌ها و عناصر قاب‌ها استفاده کرده است. در روش پیشنهادی چامبرز [15]، هر قاب به‌صورت مجموعه‌ای از رویدادهای متصل‌به‌هم در نظر گرفته شده است. برای تعیین

[1] Precision
[2] Recall
[3] https://www.wikipedia.org/

مؤلفه‌های لازم برای توصیف هر رویداد از الگوهای مبتنی بر تجزیه‌گر وابستگی و خوشه‌بندی افعال کنشی مرتبط به هم استفاده شده است.

جدول (۴) نتایج به‌دست‌آمده به‌وسیلهٔ روش پیشنهادی و سایر روش‌ها را در مسأله تشخیص قاب بر روی مجموعه‌داده آزمون نشان می‌دهد. در اینجا تنها قاب‌هایی در نظر گرفته شده است که به‌وسیلهٔ افعال کنشی قابل شناسایی هستند. در مسأله تشخیص قاب، روش پیشنهادی به کارایی ٪۴۹/۳۵ برحسب معیار F_1 دست یافته است. در مقایسه با روش‌های UTD-SRL، AFE و چامبرز، میزان بهبود روش پیشنهادی برحسب معیار F_1 به‌ترتیب برابر ٪۲/۸۵، ٪۲/۲ و ٪۰/۴۵ است. برحسب معیار دقت (P)، روش پیشنهادی در رتبه نخست قرار دارد و بهبود قابل‌توجهی نسبت به سایر روش‌ها دارد. روش چامبرز در رتبه دوم، و روش‌های AFE و UTD-SRL در رتبه‌های سوم و چهارم قرار دارند. برحسب معیار بازخوانی (R) روش چامبرز در رتبه نخست و روش پیشنهادی در رتبه دوم قرار گرفته است. نرخ یادآوری همه روش‌ها در حد مطلوب نیست که این موضوع نشان می‌دهد که مسأله استخراج قاب یک وظیفه پیچیده بوده و برای بهبود نتایج باید تلاش زیادی صورت گیرد.

(جدول-۴): مقایسه نتایج به‌دست‌آمده به‌وسیلهٔ روش پیشنهادی و سایر روش‌ها برای مسأله استخراج قاب در مجموعه‌داده فارسی

(Table-4): Comparison of results obained by the proposed method and other methods for the frame extraction problem in Farsi dataset

روش	P (%)	R (%)	F_1(%)
[19] UTD-SRL	۶۳/۲۵	۳۶/۷۷	۴۶/۵۰
[18] AFE	۶۶/۳۵	۳۶/۵۵	۴۷/۱۳
چامبرز [15]	۶۷/۵۵	**۳۸/۳۲**	۴۸/۹۰
روش پیشنهادی	**۶۹/۸۵**	۳۸/۱۵	**۴۹/۳۵**

جدول (۵) کارایی روش پیشنهادی و سایر روش‌ها را در استخراج عنوان خصلت‌های قاب‌ها بر روی مجموعه‌داده آزمون نشان می‌دهد. در تشخیص عنوان خصلت‌ها، برحسب معیار F_1، روش پیشنهادی نسبت به سایر روش‌ها بهبود را نشان می‌دهد. میزان بهبود روش پیشنهادی در مقایسه با روش‌های UTD-SRL، AFE و چامبرز برحسب معیار F_1 به‌ترتیب برابر ۴/۱۷، ۰/۵۲ و ۰/۴۷ است. برحسب معیار دقت و بازخوانی به‌ترتیب روش پیشنهادی و روش چامبرز بهترین کارایی را کسب کرده‌اند. کارایی پایین روش‌ها در جدول (۵) نشان می‌دهد که مسأله استخراج عنوان خصلت‌ها یک مسأله دشوار است.

(جدول-۵): مقایسه نتایج به‌دست‌آمده به‌وسیلهٔ روش پیشنهادی و سایر روش‌ها برای مسأله استخراج عناوین خصلت‌ها در مجموعه داده فارسی

(Table-5): Comparison of results obtained by the proposed method and other methods for the attributes' titile extraction problem in Farsi dataset

روش	P (%)	R (%)	F_1(%)
[19] UTD-SRL	۴۹/۱۱	۲۸/۵۴	۳۶/۱۰
[18] AFE	۵۸/۳۳	۳۰/۱۵	۳۹/۷۵
چامبرز [15]	۵۲/۱۹	۳۲/۱۷	۳۹/۸۰
روش پیشنهادی	۵۹/۱۷	۳۰/۵۲	۴۰/۲۷

جدول (۶) نتایج به‌دست‌آمده به‌وسیلهٔ روش‌های استخراج اطلاعات را در مسأله استخراج مقادیر خصلت‌ها برحسب معیارهای کارایی نشان می‌دهد. برحسب معیار بازخوانی، روش UTD-SRL بهترین عملکرد را نسبت به سایر روش‌ها دارد. روش پیشنهادی رتبه دوم را از نظر معیار بازخوانی کسب کرده و بدترین عملکرد مربوط به روش چامبرز است. میزان اختلاف روش پیشنهادی با روش UTD-SRL برحسب معیار بازخوانی برابر ۰/۷۳ است. برحسب معیارهای F_1 و دقت، روش پیشنهادی بهترین عملکرد را نسبت به سایر روش‌ها دارد. دلایل این موضوع را می‌توان به‌صورت زیر بیان کرد:

- مجموعه‌داده مورداستفاده شامل جملات کامل غیرساختاریافته هستند که دست‌کم شامل یک فعل کنشی اصلی هستند؛ لذا مؤلفه استخراج قاب‌های مسندآرگومان بر روی این مجموعه‌داده از کارایی بیشتری برخوردار است که در کارایی نهایی اثر مثبتی دارد. این مورد دلیل اصلی برتری کارایی روش پیشنهادی بر روی مجموعه‌داده موردنظر در مقایسه با سایر روش‌ها است.
- تحلیل هم‌ارجاعی متن و جایگزینی ضمایر با مرجع آن‌ها که در بهبود وضوح مقادیر آرگومان‌ها مؤثر است.
- تحلیل معنایی متن و اتصال مقادیر خصلت‌ها به مدخل‌های آنتولوژی BabelNet. در مرحله تحلیل معنایی متن، عبارات چندکلمه‌ای که به‌عنوان مقادیر خصلت‌ها هستند به‌درستی مشخص می‌شوند که این موضوع موجب بالا رفتن دقت روش پیشنهادی می‌شود.

(جدول-۶): مقایسه نتایج روش پیشنهادی و سایر روش‌ها در مسأله استخراج مقادیر خصلت‌های قاب در مجموعه‌داده فارسی

(Table-6): Comparison results of the proposed method and other methods for the attributes' values extraction problem in Farsi dataset

روش	P (%)	R (%)	F_1(%)
[19] UTD-SRL	۴۹/۶۷	۲۳/۱۸	۳۱/۶۱
[18] AFE	۵۲/۱۹	۲۰/۱۵	۲۹/۰۷
چامبرز [15]	۵۴/۲۷	۱۹/۶۹	۲۸/۹۰
روش پیشنهادی	۵۵/۳۱	۲۲/۴۵	۳۱/۹۴

با توجه به نتایج به‌دست‌آمده در جدول‌های (۴ تا ۶)، واضح است که روش پیشنهادی در مقایسه با روش‌های دیگر، کارایی بیشتری در استخراج اطلاعات ساختاریافته دارد. روش‌های دیگر فقط به استخراج اطلاعات ساختاریافته تمرکز کرده و به معنای اطلاعات توجهی نکرده‌اند. این در حالی است که روش پیشنهادی علاوه بر استخراج اطلاعات ساختاریافته به موضوع نگاشت اطلاعات به مدخل‌های آنتولوژی‌های عمومی نیز توجه کرده است. این کار موجب تسهیل پردازش متون زبان فارسی می‌شود؛ زیرا به‌سادگی می‌توان اطلاعات استخراج‌شده را به متون سایر زبان‌ها ترجمه و از ابزارهای سایر زبان‌ها برای پردازش اطلاعات استفاده کرد. در حالت کلی، کارایی روش‌های موردبحث بسیار مطلوب نیست و تلاش زیادی برای بهبود نتایج نیاز است. این موضوع دشواری مسأله استخراج اطلاعات ساختاریافته از متون غیرساختاریافته زبان فارسی را نشان می‌دهد. باید به این نکته اذعان کرد که در متون ساختاریافته و یا متونی که دارای جملات کوتاه باشند و یا جملاتی که دارای افعال کنشی نیستند، قطعاً مؤلفه استخراج قاب‌های مسندآرگومان کارایی مطلوبی نخواهد داشت و کارایی سیستم کاهش خواهد یافت. بنابراین روش پیشنهادی بر روی مجموعه‌داده‌هایی که شامل جملاتی هستند که دارای افعال کنشی هستند، کارایی مطلوبی دارد. با توجه به ساختار روش پیشنهادی و استفاده از آنتولوژی‌های موجود برای پردازش متن، روش پیشنهادی می‌تواند بر روی متون سایر زبان‌ها نیز اعمال شود، به‌شرط آنکه از ابزارهای پیش‌پردازش متناسب با آن زبان‌ها نیز استفاده شود.

به‌منظور بررسی بیشتر، روش پیشنهادی بر روی یک مجموعه‌داده انگلیسی مورد ارزیابی قرار گرفته است. این مجموعه‌داده شامل یکصد جمله کامل بوده که در آن هر جمله حداقل دارای یک فعل کنشی است. مجموعه داده مورد استفاده از نشانی زیر قابل دسترسی است:

https://www.ubonab.ac.ir/uploads/user/310/files/engdata.txt

(جدول-۷): مقایسه نتایج به‌دست‌آمده توسط روش پیشنهادی و سایر روش‌ها بر روی مجموعه‌داده انگلیسی

(Table-7): Comparison of results obained by the proposed method and other methods on englisgh dataset

مسأله	روش	*P (%)*	*R (%)*	*F_1(%)*
استخراج قاب	[19] UTD-SRL	۶۶/۹۷	۳۸/۲۶	۴۸/۷۰
	[18] AFE	۶۹/۱۷	۳۹/۹۳	۵۰/۶۳
	چامبرز [15]	**۷۵/۱۱**	**۴۳/۹۶**	**۵۵/۴۶**
	روش پیشنهادی	۷۳/۶۵	۴۳/۰۲	۵۴/۳۱
استخراج عناوین خصلت‌ها	[19] UTD-SRL	۶۱/۳۰	۲۹/۸۳	۴۰/۱۳
	[18] AFE	۶۰/۸۷	۳۰/۴۷	۴۰/۶۱
	چامبرز [15]	**۶۵/۹۳**	۳۰/۹۴	**۴۲/۱۲**
	روش پیشنهادی	۶۱/۲۵	**۳۱/۵۵**	۴۱/۶۵
استخراج مقادیر خصلت‌ها	[19] UTD-SRL	۵۴/۶۹	۲۳/۹۲	۳۳/۲۸
	[18] AFE	۵۹/۷۰	۲۴/۳۰	۳۴/۵۴
	چامبرز [15]	۵۵/۹۹	**۲۵/۴۴**	۳۴/۹۸
	روش پیشنهادی	**۶۰/۱۵**	۲۵/۰۲	**۳۵/۳۴**

جدول (۷) کارایی روش پیشنهادی و سایر روش‌ها در استخراج قاب‌ها، عناوین و مقادیر خصلت‌ها بر روی مجموعه‌داده انگلیسی را نشان می‌دهد. برای ارزیابی روش پیشنهادی بر روی مجموعه داده انگلیسی، تغییراتی در ابزارهای پیش‌پردازش صورت گرفته است تا برای پردازش زبان انگلیسی مناسب باشند. به همین منظور، از ابزارهای تحلیل هم‌ارجاعی، جمله‌یاب و تجزیه وابستگی که در بسته CoreNLP[1] فراهم آمده، استفاده شده است. برای استخراج گزاره‌های مسندآرگومان، ابزار SENNA[2] مورد

[1] https://stanfordnlp.github.io/CoreNLP/
[2] http://ronan.collobert.com/senna/

استفاده قرار گرفته است. همچنین از منابع پردازش زبانی SemLink[1] و وردنت[2] برای نگاشت گزاره‌های مسند آرگومان به نقش‌های محدود به دامنه استفاده شده است. منبع نگاشت SemLink حاوی اطلاعات مناسبی برای نگاشت نقش‌های معنایی بین PropBank، VerbNet و FrameNet است.[29]

همان‌گونه که در جدول (۷) آمده است، در مسأله تشخیص قاب و عناوین خصلت‌ها، روش چامبرز بهترین عملکرد را در مقایسه با سایر روش‌ها داشته است. روش پیشنهادی بر روی مسأله تشخیص قاب و عناوین خصلت‌ها، جایگاه دوم و روش AFE جایگاه سوم را کسب کرده است. در مسأله استخراج مقادیر خصلت‌ها، روش پیشنهادی بهترین کارایی را بر حسب معیارهای دقت و F_1 کسب کرده است. روش چامبرز نیز برحسب معیار بازخوانی رتبه نخست را دارد. نتایج نشان می‌دهد که روش پیشنهادی، در مقایسه با سایر روش‌های همتا، کارایی برابر و یا کارایی نزدیک به آن‌ها را کسب کرده است. با این حال، نتایج با میزان بهینه فاصله زیادی دارند و تلاش زیادی برای بهبود روش‌های استخراج اطلاعات ساختاریافته در هر دو زبان فارسی و انگلیسی موردنیاز است.

۵- نتیجه‌گیری

در این مقاله، مسأله استخراج اطلاعات ساختاریافته معنایی از متون غیرساختاریافته وب بررسی شده و روشی معنایی مبتنی بر ترکیب ابزارها و منابع پردازش زبانی ارائه شد. اطلاعات استخراج‌شده علاوه‌بر ساختاریافتگی به آنتولوژی‌های عمومی BabelNet و DBPedia متصل شده‌اند. این کار موجب تسهیل در پردازش و ترجمه ماشینی متون فارسی می‌شود. نتایج آزمایش‌ها نشان می‌دهد که روش پیشنهادی از کارایی مطلوبی در مقایسه با سایر روش‌ها برخوردار است. یکی از کارهای آینده، بهبود ابزارهای پیش‌پردازش و مؤلفه‌های تحلیل معنایی متن به‌ویژه مؤلفه رفع ابهام معنایی کلمات و استخراج قاب‌های مسندآرگومان است که می‌تواند کارایی روش پیشنهادی را بهبود دهد؛ همچنین بسط منبع نگاشت نقش‌ها نیز باید موردتوجه قرار گیرد که با این کار می‌توان محدودیت روش پیشنهادی در نگاشت قاب‌های مسند آرگومان مستقل از دامنه به اطلاعات ساختاریافته محدود به دامنه را رفع کرد. ارزیابی روش پیشنهادی بر روی متون چندزبانی و نیز استفاده از مجموعه داده بزرگ‌تر به‌عنوان یکی دیگر از کارهای آینده در نظر گرفته شده است.

[1] https://verbs.colorado.edu/semlink/
[2] https://wordnet.princeton.edu/

۶- مراجع 6- References

[1] A. A. Barforoush, H. Shirazi, and H. Emami, "A new classification framework to evaluate the entity profiling on the Web: past, present and future," *ACM Comput. Surv.*, vol. 50, no. 3, pp. 1–39, 2017.

[2] H. Emami, H. Shirazi, and A. A. Barforoush, "A Semantic approach to person profile extraction from Farsi documents," *J. Inf. Syst. Telecommun.*, vol. 4, no. 4, pp. 232–243, 2016.

[3] W. Li, R. Srihari, C. Niu, and X. Li, "Entity profile extraction from large corpora," in *Pacific Association for Computational Linguistics Conference (PACLING-2003), Harifax, Canada*, 2003.

[4] Y. Chen, S. Y. Mei Lee, and C. R. Huang, "A robust web personal name information extraction system," *Expert Syst. Appl.*, vol. 39, no. 3, pp. 2690–2699, 2012.

[5] U. Distant and S. Machine, "Domain-Specific Relation Extraction Using Distant Supervision Machine Learning," in *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015)*, 2015, pp. 978–989.

[6] N. T. Nakashole, "Automatic extraction of facts, relations, and entities for web-scale knowledge base population," University of Saarland, 2013.

[7] S. Soderland, B. Roof, B. Qin, and S. Xu, "Adapting Open Information Extraction to Domain-Specific Relations," *AI Mag.*, vol. 31, no. 3, pp. 93–102, 2010.

[8] S. Soderland, J. Gilmer, R. Bart, O. Etzioni, and D. Weld, "Open Information Extraction to KBP Relations in 3 Hours," in *Proceedings of TAC-KBP 2013, Maryland, USA*, 2013.

[9] M. Shamsfard, "Challenges and open problems in Persian text processing," in *Proceedings of 5th Language & Technology Conference (LTC), Poznań, Poland*, 2011, pp. 65–69.

[10] Y. Chen, S. Lee, and C. Huang, "Polyuhk: A robust information extraction system for web personal names," in 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, Madrid, Spain, 2009.

[11] S. Lyons and D. Smith, "Domain-specific information extraction structures," *Proc. - Int. Work. Database Expert Syst. Appl. DEXA*, vol. 2002–Janua, pp. 80–84, 2002.

[12] Y. Shinyama and S. Sekine, "Preemptive information extraction using unrestricted relation discovery," in *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, New*

York, USA, 2006, pp. 304–311.

[13] N. Chambers and D. Jurafsky, "Unsupervised learning of narrative event chains," in *Proceedings of the Association of Computational Linguistics (ACL), Columbus, Ohio*, 2008, pp. 789–797.

[14] N. Kasch and T. Oates, "Mining script-like structures from the web," in *Proceedings of the NAACL HLT, Los Angeles, California*, 2010, pp. 34–42.

[15] N. Chambers and D. Jurafsky, "Template-Based Information Extraction without the Templates," in *HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Stroudsburg, PA, USA*, 2011.

[16] M. Banko, M. Cafarella, and S. Soderland, "Open information extraction from the web," in *International Joint Conferences on Artificial Intelligence, Hyderabad, India*, 2007, pp. 2670–2676.

[17] R. Johansson and P. Nugues, "LTH : Semantic Structure Extraction using Nonprojective Dependency Trees," in *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague*, 2007, pp. 227–230.

[18] M. Scaiano and D. Inkpen, "Automatic frame extraction from sentences," in *Canadian Conference on Artificial Intelligence, Kelowna, British Columbia*, 2009, pp. 110–120.

[19] C. Bejan, C.A., Hathaway, "UTD-SRL: A Pipeline Architecture for Extracting Frame Semantic Structures," in *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic*, 2007.

[20] H. Fadaei and M. Shamsfard, "Extracting conceptual relations from Persian resources," in *ITNG2010 - 7th International Conference on Information Technology: New Generations, Las Vegas, Nevada, USA*, 2010, pp. 244–248.

[21] M. Moradi, B. Vazirnezhad, and M. Bahrani, "Commonsense Knowledge Extraction for Persian Language: A Combinatory Approach," *Iran. J. Inf. Process. Manag.*, vol. 31, no. 1, pp. 109–124, 2015.

[22] M. Shamsfard, "Lexico-syntactic and Semantic Patterns for Extracting Knowledge from Persian Texts," *Int. J. Comput. Sci. Eng.*, vol. 2, no. 6, pp. 2190–2196, 2010.

[23] H. Emami, H. Shirazi, A. A. Barforoush, and M. Hourali, "A Pattern-Matching Method for Extracting Personal Information in Farsi Content," *U.P.B. Sci. Bull., Ser. C*, vol. 78, no. 1, pp. 125–138, 2016.

[24] R. Al-Rfou, V. Kulkarni, B. Perozzi, and S. Skiena, "Polyglot-NER: Massive Multilingual Named Entity Recognition," in *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada*, 2015, pp. 586–594.

[25] F. Fallahi and M. Shamsfard, "Recognizing Anaphora Reference in Persian Sentences," *Int. J. Comput. Sci. Issues*, vol. 8, no. 2, pp. 324–329, 2011.

[26] A. Moro, A. Raganato, and R. Navigli, "Entity linking meets word sense disambiguation: a unified approach," *Trans. Assoc. Comput. Linguist.*, vol. 2, pp. 231–244, 2014.

[27] Z. M. Arani and A. Abdollahzadeh Barforoush, "Semantic Role Labeling using Syntactic Dependency Analysis and Noun Semantic Catergory," in *20th Annual Conference of Computer Society of Iran, Mashhad, Iran (In Farsi)*, 2015, pp. 619–624.

[28] K. Kipper, A. Korhonen, N. Ryant, and M. Palmer, "A large-scale classification of English verbs," *Lang. Resour. Eval.*, vol. 42, no. 1, pp. 21–40, 2008.

[29] E. Loper, S. Yi, and M. Palmer, "Combining lexical resources: mapping between propbank and verbnet," in Proceedings of the 7th International Workshop on Computational Linguistics, Tilburg, Netherlands, 2007.

[30] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

[31] C. Baker and M. Ellsworth, "SemEval'07 Task 19: Frame Semantic Structure Extraction," in *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic*, 2007, pp. 99–104.

حجت امامی در رشته مهندسی کامپیوتر گرایش هوش مصنوعی فارغ التحصیل شده است. ایشان دانشیار گروه مهندسی کامپیوتر دانشگاه بناب است. زمینه‌های پژوهشی وی شامل داده‌کاوی، یادگیری ماشین، سیستم‌های چندعاملی، الگوریتم‌های اکتشافی و فرااکتشافی، هوش جمعی و کاربردهای آن است. هم‌اکنون او در زمینه استفاده از یادگیری ماشین در حوزه مدیریت ترافیک هوایی، تشخیص بیماری در حوزه پزشکی و جایابی گره‌ها در شبکه‌های نوری کار پژوهشی خود را ادامه می‌دهد.

نشانی رایانامه ایشان عبارت است از:

emami@ubonab.ac.ir