Persian Causality Corpus (PerCause) and the Causality Detection Benchmark

Zeinab Rahimi
PhD Candidate in Computer Engineering; NLP Research Laboratory; Shahid Beheshti University; Tehran, Iran;
Email: rahimi.zeinab@gmail.com

Mehrnoush ShamsFard*
PhD in Computer Engineering; Associate Professor;
NLP Research Laboratory; Shahid Beheshti University; Tehran, Iran;
Email: m-shams@sbu.ac.ir

Iranian Journal of Information Processing and Management

Received: 03, Jan. 2022 **Accepted: 12, Apr. 2022**

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 38 | No. 2 | pp. 607-638
Winter 2023
https://doi.org/10.35050/JIPM010.2022.036

Abstract: Recognizing causal elements and causal relations in a text is among the challenging issues in natural language processing (NLP), specifically in low-resource languages such as Persian. In this research, we prepare a causality human-annotated corpus for the Persian language. This corpus consists of 4446 sentences and 5128 causal relations. Three labels of Cause, Effect, and Causal mark are specified to each relation, if possible. We used this corpus to train a system for detecting causal elements' boundaries.

Also, we present a causality detection benchmark for three machine-learning methods and two deep learning systems based on this corpus. Performance evaluations indicate that our best total result is obtained through the CRF classifier, which provides an F-measure of 0.76. In addition, the best accuracy (91.4%) is obtained through the BiLSTM-CRF deep learning method.

Keywords: PerCause; Causality Annotated Corpus; Causality Detection; Deep Learning; CRF

* Corresponding Author

معرفی و آزمون پیکرهٔ علیت PerCause برای شناسایی روابط علّی فارسی

زینب رحیمی
دانشجوی دکتری مهندسی کامپیوتر؛ آزمایشگاه پردازش زبان طبیعی؛ دانشگاه شهید بهشتی؛ تهران، ایران؛
rahimi.zeinab@gmail.com

مهرنوش شمس‌فرد
دکتری مهندسی کامپیوتر؛ دانشیار؛ آزمایشگاه پردازش زبان طبیعی؛ دانشگاه شهید بهشتی؛ تهران، ایران؛
پدیدآور رابط m-shams@sbu.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۴۰۰/۱۰/۱۳ | پذیرش: ۱۴۰۱/۰۱/۲۳ | مقاله برای اصلاح به مدت ۱۶ روز نزد پدیدآوران بوده است.

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران
(ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۸ | شمارهٔ ۲ | ص‌ص ۶۰۷-۶۳۸
زمستان ۱۴۰۱
https://doi.org/10.35050/JIPM010.2022.036

چکیده: شناسایی روابط علّی و تعیین مرز عناصر علّی در متن از جمله مسائل چالش‌برانگیز در پردازش زبان طبیعی، به‌ویژه در زبان‌های کم‌منبع مانند زبان فارسی است. در این پژوهش در راستای آموزش سیستمی برای شناسایی روابط علّی و مرز عناصر آن، یک پیکرهٔ علّیت برچسب‌خوردهٔ انسانی برای زبان فارسی معرفی می‌شود. این مجموعه شامل ۴۴۴۶ جمله (مستخرج از پیکرهٔ «بیجن‌خان» و متن یک سری کتاب) و ۵۱۲۸ رابطهٔ علّی است و در صورت وجود، سه برچسب علت، معلول، و نشانهٔ علّی برای هر رابطه مشخص شده است. در این پژوهش از پیکره برای آموزش سیستمی به‌منظور تشخیص مرزهای عناصر علّی استفاده شده است. همچنین، یک بستر آزمون شناسایی علّیت با سه روش یادگیری ماشین و دو سیستم یادگیری عمیق مبتنی بر این پیکره ارائه شده است. ارزیابی‌های عملکرد نشان می‌دهد که بهترین نتیجهٔ کلی از طریق طبقه‌بندی‌کنندهٔ CRF به‌دست می‌آید که معیار F برابر ۷۶ درصد را ارائه می‌کند. افزون بر این، بهترین صحت (۹۱/۴ درصد) در روش یادگیری عمیق BiLSTM-CRF به‌دست آمده است. به نظر می‌رسد که وجود CRF به‌دلیل مدل‌سازی بافتار به بهبود دقت سیستم منجر می‌شود.

کلیدواژه‌ها: PerCause، شناسایی عبارات علّی، CRF، یادگیری عمیق

۱. مقدمه

استخراج روابط علّی یک معضل آشکار در پردازش زبان طبیعی است که به‌طور عمده شامل استفاده از تحلیل معنایی است. این رویکرد در بسیاری از وظایف پردازش زبان طبیعی[1] مانند تشخیص استلزامات متنی، سیستم پرسش و پاسخ، پیش‌بینی رویدادها و استخراج روایت استفاده می‌شود. منظور از رابطهٔ علّی در این پژوهش رابطه بین جملات و عباراتی است که نشانگر وقوع واقعه یا وجود حالتی بوده و علت و معلول یکدیگر هستند. برای روشن شدن این مسئله به مثال‌های زیر توجه کنید:

(۱) کیومرث مسموم شده است، زیرا سیب سمی را خورد.

(۲) باران می‌بارد. خیابان‌ها لغزنده هستند.

(۳) جاده‌های باریک خطرناک هستند.

(۴) سفیدبرفی پس از خوردن سیب سمی مسموم شد.

(۵) بچه‌ای که با کبریت بازی می‌کرد، خانه را سوزاند.

همهٔ این موارد شامل رابطهٔ علّی هستند. یک رابطهٔ علّی به‌طور معمول، از سه عنصر علت، معلول و نشانهٔ علّیت تشکیل شده است. به‌عنوان نمونه، در مثال (۱)، «کیومرث مسموم شده است»، معلولِ «سیب سمی را خورد» است و «زیرا» نشانهٔ علّیت است. همان‌طور که در مثال‌های بالا مشاهده می‌شود، عناصر علّی ممکن است کلمات، عبارات یا گزاره باشند و در یک یا دو جمله رخ دهند.

علّیت در شاخه‌های مختلف دانش از قبیل روان‌شناسی، زبان‌شناسی، فلسفه و علم کامپیوتر مطالعه شده است. روابط علّی از دیدگاه‌های مختلفی دسته‌بندی می‌شوند؛ برای مثال، بر اساس ظهور یا عدم ظهور نشانهٔ علّیت در جمله، نشانهٔ علّیت ممکن است در جمله ظاهر شود یا نشود. همچنین، در صورت ظهور هم می‌تواند مبهم باشد یا نباشد. به‌عنوان نمونه، در مثال (۳)، «خطر» معلولِ «جاده‌های باریک» است، اما هیچ نشانهٔ علّیت صریحی نداریم. یا در مثال (۴)، علامت علّی «پس» مبهم است، زیرا این کلمه همیشه نشانگر علّیت نیست.

از سوی دیگر، ساختارهای علّی ممکن است صریح یا ضمنی باشند (Girju 2003).

1. natural language processing (NLP)

بنابراین، ممکن است علت یا معلول به‌طور ضمنی در یک جملهٔ علّی بیان شوند. به‌عنوان نمونه، در مثال (۵)، «بازی کردن بچه با کبریت» دلیل مستقیم سوختن خانه نیست. بلکه، آتش دلیل سوختن است. بنابراین، یافتن مرزهای دقیق روابط علّی امری مهم و چالش‌برانگیز است و ممکن است نیازمند دانش عرفی باشد.

از نقطه‌نظر دیگر، شرط، عواقب، و دلیل زیرمجموعه‌های رابطهٔ علّی هستند. گاهی روابط علّی را در قالب شرط بیان می‌کنیم. این مسئله زمانی است که علت فرضی باشد (اگر زیبا بود، ازدواج می‌کرد). گاهی رابطهٔ علّی را در قالب رویدادهایی که بعد از یک حالت یا رویداد اتفاق می‌افتند، در نظر می‌گیریم. این مسئله زمانی است که تأثیر غیرمستقیم یا غیرعمد در میان باشد (استعفای او باعث حسرت همه شد) و دستهٔ سوم زمانی است که علّیت در مورد تصمیم، باور، احساس یا عمل باشد (Blanco, Castell and Moldovan 2008)؛ مانند (من رفتم، چون فکر کردم جالب است).

عناصر علّی ممکن است با استفاده از روش‌های مبتنی بر قانون یا روش‌های یادگیری ماشین که به پیکرهٔ برچسب‌گذاری‌شده نیاز دارند، شناسایی شوند. به‌طور کلی، به نظر می‌رسد که رویکردهای مبتنی بر پیکره برای چنین مواردی مناسب هستند (Goyal, Manish and Vishal 2017). سیستم‌های تشخیص علّیت مبتنی بر پیکره برای تعیین مرزهای هر عنصر علّی در یک جمله یا یک جفت جمله و سپس طبقه‌بندی آن‌ها به دسته‌های از پیش تعریف‌شدهٔ علت، معلول و نشانهٔ علّیت استفاده می‌شوند. دادگان آموزشی علّی در حال حاضر در زبان‌هایی مانند انگلیسی وجود دارند (که برخی از آن‌ها در بخش ۲ معرفی شده‌اند)، با این حال، تا جایی که می‌دانیم، چنین مجموعه‌ای برای زبان فارسی ساخته نشده است. این مقاله یک پیکرهٔ برچسب‌خورده توسط انسان برای علّیت به زبان فارسی (PerCause) و بستر آزمونی با ارائهٔ مدل‌های آموزش‌یافته بر اساس مجموعهٔ دادهٔ ایجادشده ارائه می‌کند. PerCause انواع ساده و عام علّیت را پوشش می‌دهد و موارد پیچیده مانند علّیت‌های متافیزیکی، ضعیف، نفی یا تودرتو در حوزهٔ کار این پژوهش نیست.

ادامهٔ این مقاله به شرح زیر سازماندهی شده است. بخش ۲، پیشینهٔ پژوهش را بررسی می‌کند. بخش ۳، به معرفی روش پژوهش می‌پردازد. این بخش شامل دو زیر بخش است: ابتدا پیکرهٔ برچسب‌خوردهٔ انسانی علّیت فارسی (PerCause) معرفی شده و روش تولید آن به تفصیل بیان می‌گردد. سپس، روش‌های مختلف یادگیری ماشین و یادگیری عمیق را

برای استخراج روابط علّی بر روی PerCause معرفی می‌کند. در بخش ۴، تجزیه و تحلیل یافته‌ها و جزئیات ارزیابی‌ها بیان می‌گردد. سرانجام، در بخش ۵، بحث مقاله انجام شده و نتیجه‌گیری صورت می‌گیرد.

۲. پیشینهٔ پژوهش

در این بخش، پژوهش‌های مرتبط را در دو دستهٔ اصلی مرور می‌کنیم: ۱) پژوهش‌هایی که به معرفی روش‌های شناسایی علّیت پرداخته‌اند و ممکن است به توسعهٔ نیمه‌خودکار یک پیکره منجر شود، و ۲) پژوهش‌هایی که به معرفی پیکره‌های برچسب‌خوردهٔ انسانی برای علّیت پرداخته‌اند.

۲-۱. پژوهش‌های حوزهٔ شناسایی علّیت

استخراج رابطهٔ علّی به‌طور معمول، از طریق دو رویکرد انجام می‌شود: روش‌های مبتنی بر قانون و روش‌های یادگیری ماشین. بعضی پژوهشگران برای شناسایی روابط علّی صریح، الگوهای زبانی را به‌کار می‌برند؛ برای مثال، Garcia (1997) از یک مدل معنایی استفاده کرده که الگوهای کلامی علّی را به‌عنوان شاخص‌های زبانی مانند (NP۱ فعل علّی NP2) برای تشخیص روابط علّی در متون فرانسوی طبقه‌بندی می‌کند. Khoo, Syin and Yun (2000) برای استخراج اطلاعات علت-معلولی از متون روزنامه‌های تجاری و پزشکی از الگوهای کلامی-زبانی از پیش تعریف شده استفاده کردند. Luo et al. (2016) در یک مطالعهٔ دیگر چارچوبی را پیشنهاد کردند که به‌طور خودکار شبکه‌ای از اصطلاحات علّی را با استفاده از الگوهای مشخص (مثلاً A منجر به B می‌شود یا اگر A سپس B) از یک مجموعهٔ وب بزرگ جمع آوری می‌کند. آن‌ها بر اساس این شبکه معیاری را برای مدل‌سازی قدرت علّیت بین عبارت‌ها پیشنهاد کردند. در این روش روابط علّی در متون کوتاه با استفاده از منبع ایجادشده و معیارهای آماری ارائه‌شده شناسایی می‌شوند. بیان شده که این رویکرد با حاشیه‌های قابل توجهی از تمام نتایج گزارش‌شدهٔ قبلی در کار استاندارد «کوپا» درکارگاه «سِم‌ایوَل»[1] بهتر عمل می‌کند.

استخراج علّیت با یادگیری ماشین نیازمند یک مجموعه علّیت است که به‌طور عام،

1. The COPA evaluation was accepted as Task 7 of the 6th International Workshop on Semantic Evaluation (SEMEVAL)

به‌صورت دستی یا نیمه‌خودکار بر اساس الگوهای خاص تهیه می‌شود. برای نمونه در پژوهش Girju (2003)، یک مجموعهٔ آموزشی متشکل از جملات با ۶۰ فعل علّی ساده با استفاده از متون کنفرانس تِرِک-۹[1] بخش لس‌آنجلس تایمز[2] ایجاد می‌شود. او با استفاده از تجزیه‌کنندهٔ نحوی ۶۵۲۳ رابطه به شکل NP1-Verb-NP2 پیدا کرد که از این میان، ۲۱۰۱ رابطه به‌صورت دستی به‌عنوان روابط علّی و ۴۴۲۲ رابطه به‌عنوان غیرعلّی برچسب‌گذاری شدند. این روابط نمونه‌های مثبت و منفی را برای آموزش طبقه‌بندی درخت تصمیم برای طبقه‌بندی روابط علّی و غیرعلّی در متن ایجاد می‌کنند. بیان شده که دقت الگوهای استخراج شده در این پژوهش ۷۳ و بازخوانی آن‌ها ۸۸ است.

«چانگ و چوی» یک مکانیزم راه‌حل جایگزین برای یادگیری عبارات نشانه و احتمالات جفت واژگانی از پیکرهٔ خام و مستقل از دامنه به شیوه‌ای بدون نظارت پیشنهاد کردند. آن‌ها از این احتمالات برای استخراج علّیت بین اسمی و بین جمله‌ای استفاده کردند. استخراج رابطهٔ علّی بین اسم‌ها دقت ۸۱/۲۹ درصد را نشان می‌دهد که ۷/۰۵ درصد نسبت به مدل پایه بهبود یافته است (Chang and Choi 2004). در پژوهش Blanco, Castell and Moldovan (2008) مؤلفان روابط علّی صریح را که با الگوی عبارت فعلی-ارتباط دهنده-علت[3] بیان می‌شد، {relator ∈ {because, since, after, as (ارتباط‌دهنده از مجموعه خاصی انتخاب شده است) در نظر گرفته‌اند. آن‌ها از مجموعهٔ SemCor 2.1 با برچسب‌گذاری معنایی برای آموزش طبقه‌بندی‌کنندهٔ درخت تصمیم C4.5 استفاده کرده و سرانجام، به معیار F ۸۹ دست یافتند.

Mirza (2014) دستورالعمل‌های برچسب‌گذاری را بر اساس تعریف دستورالعمل برچسب‌زنی TimeML[4] از رویدادها، برای یافتن علّیت بین رویدادها پیشنهاد کرد. این دستورالعمل‌ها همهٔ انواع عمل‌ها (ثابت و مدت‌دار) و حالت‌ها را به‌عنوان رویداد در نظر می‌گیرند. او برچسب <CLINK> را برای مشخص کردن یک پیوند علّی و مفهوم سیگنال‌های علّی را با برچسب <C-SIGNAL> ایجاد کرد. Mirza and Tonelli (2016) نیز سیستم «کاتنا»[5] را برای استخراج و طبقه‌بندی رابطهٔ زمانی و علّی از متن انگلیسی ارائه کردند. در این راستا آن‌ها از تعامل بین روابط زمانی و علّی بهره‌برداری کرده و از یک رویکرد ترکیبی

1. TREC-9: Text REtrieval Conference (TREC)
2. LA TIMES
3. VerbPhrase-Relator-Cause
4. TimeML Annotation Guidelines
5. Causal and temporal relation extraction from natural language texts (CATENA)

نیـز اسـتفاده کردنـد کـه طبقه‌بندی‌کننده‌هـای مبتنـی بـر قانـون و بانظـارت را بـرای شناسـایی روابـط علّـی ترکیـب می‌کنـد. آن‌هـا سـرانجام، مفهـوم علّیـت را کـه در دسـتورالعمل‌های برچسـب‌گذاری (2014) Mirza و Mirza and Tonelli (2016) برچسـبِ CausalTimeBank پیشـنهاد شـده، پذیرفتنـد. ایـن دسـتورالعمل شـامل پدیده‌هـای CAUSE، ENABLE و PREVENT اسـت. هـدف مؤلفـان ایـن اسـت کـه وقتـی رابطـهٔ علّـی بـا معلـول، لینـک و افعـال علّـی (افعـال نـوع CAUSE، ENABLE و PREVENT) بیـان می‌شـوند، یـا هنگامـی کـه یـک نشـانهٔ علّـی رابطـهٔ علّـی را نشـان می‌دهـد، لینکـی بیـن جفـت رویـداد اختصـاص دهنـد. ارزیابی‌هـا نشـان داده کـه نتایـج امیدوارکننـده‌ای به‌دسـت آمـده و وابسـتگی ابعـاد زمانـی و علّـی تأییـد شـده‌اند.

بـه عقیـدهٔ «نینـگ» و همـکاران، چـون هـر علـت بایـد زودتـر از معلـول خـود رخ دهـد، روابـط زمانـی و علّـی ارتبـاط نزدیکـی بـا هـم دارنـد. بنابرایـن، آن‌هـا بـا اسـتفاده از مدل‌هـای شـرطی محـدود(CCM)[1] یـک چارچـوب اسـتنتاج مشـترک بـرای آن‌هـا ارائـه کردنـد. در واقـع، آن‌هـا مسـئلهٔ مشـترک را به‌عنـوان یـک مسـئلهٔ برنامه‌ریـزی خطـی عـدد صحیـح (ILP)[2] فرمولـه کـرده و محدودیت‌هـای ذاتـی ماننـد زمـان و علّیـت را تقویـت کردنـد. آن‌هـا سـرانجام، عنـوان کردنـد کـه چارچـوب اسـتنتاج مشـترک منجـر بـه بهبـود آمـاری معنـادار در اسـتخراج روابـط زمانـی و علّـی از متـن می‌شـود (Ning et al. 2019).

در پژوهـش Dasgupta et al. (2018)، یـک معمـاری شـبکهٔ عصبـی بازگشـتی بـا اطلاعـات زبانـی (یـک مـدل حافظـهٔ کوتاه‌مـدت طولانـی[3] دوطرفـه کـه بـا یـک لایـهٔ زبانـی اضافـی تقویـت شـده اسـت)، بـرای اسـتخراج خـودکار روابـط علـت و معلولـی از متـن پیشـنهاد شـده اسـت. در معمـاری پیشـنهادی، از تعبیه‌هـای در سـطح کلمـه و چنـد ویژگـی زبانـی دیگـر بـرای تشـخیص رویدادهـای علّـی و معلول‌هـای آن‌هـا (کـه در یـک جملـه ذکـر شـده اسـت) اسـتفاده می‌شـود.

سـرانجام، در پژوهـش «هاشـیموتو» روشـی بـا نظـارت ضعیـف بـرای اسـتخراج دانـش علّـی ماننـد Protectionism -> Trade war از مقـالات «ویکی‌پدیـا» بـه چندیـن زبـان ارائـه شـد. ایـدهٔ کلیـدی ایـن مطالعـه اسـتفاده از بخش‌هـای توصیف‌کننـدهٔ علّیـت و چندزبانـه بـودن «ویکی‌پدیـا» بـود. بیـان شـده کـه ایـن روش بـه دقـت ۹۸ و بازخوانـی ۶۴ رسـیده و بـرای علـت و معلول‌هـای بافاصلـه در متـن، عملکـرد موفقـی داشـته اسـت (Hashimoto 2019).

1. constrained conditional model (CCM)
2. integer linear programming (ILP)
3. long-short term memory (LSTM)

۲-۲. پژوهش‌هایی که به معرفی پیکره‌های برچسب‌خوردهٔ انسانی برای علّیت پرداخته‌اند

چندین منبع برچسب‌گذاری‌شدهٔ دستی برای علّیت وجود دارد. منابع فعلی مانند VerbNet که در مقالهٔ Schuler (2005) و PropBank که در پژوهش Palmer et al. (2005) معرفی شده‌اند، شامل افعال دارای بار علّیت است. همین‌طور Schneider et al. (2016) نیز شمای حرف اضافه را موارد علّی معرفی می‌کنند؛ هرچند این موارد فقط کلاس‌های خاصی از کلمات را پوشش می‌دهند.

تا جایی که می‌دانیم، هیچ مجموعهٔ برچسب‌گذاری‌شدهٔ علّیت برای زبان فارسی وجود ندارد. اما چند مجموعه داده علّیت برچسب‌گذاری‌شده به‌صورت دستی برای زبان انگلیسی وجود دارد. برای نمونه، BECAUSE (Dunietz, Levin and Carbonell 2015)، BioCause (Mihăilă et al. 2013)، و CaTeRS (Mostafazadeh et al. 2016; Dunietz, Levin and Carbonell 2015).

«دانیتز، لوین و کاربونلی» در خصوص BECAUSE سه نوع علّیت را شناسایی می‌کنند که هر کدام معنای اندک متفاوتی دارد. اولین علّیت «نتیجه» است که در آن علت به‌طور طبیعی به معلول منتهی می‌شود (به‌عنوان نمونه، «ما به‌دلیل مقررات ناکافی، مشکل اقتصادی جدی داریم»). علّیت دوم، «انگیزه» است که در آن عاملی علت را درک می‌کند. بنابراین، آگاهانه فکر می‌کند، احساس می‌کند، یا چیزی را انتخاب می‌کند (برای مثال، «زمان زیادی نداریم؛ پس بیایید سریع حرکت کنیم»). در نهایت، در علّیت نوع «هدف»، یک عامل برای درست جلوه دادن علت معلول را انتخاب می‌کند. برای مثال، (آن‌ها را در رسیدگی به شکایات راهنمایی کنید تا بتوانند مشکلات را حل کنند) (Dunietz, Levin and Carbonell 2015).

«مصطفی‌زاده» و همکاران در خصوص CaTeRS، بر ۹ رابطهٔ علّی همراه با رابطهٔ زمانی تمرکز می‌کنند؛ از جمله: «علت» (قبل از/ همپوشانی)، «فعال‌سازی» (قبل/ همزمان)، «جلوگیری» (قبل/ همزمان) و «علت به پایان» (قبل/ همزمان/ در طول) (Mostafazadeh et al. 2016). در واقع، CaTeRS چهار رابطهٔ علّی علت، فعال‌سازی، جلوگیری، و علت به پایان را در کنار فرض «A علت/ فعال کردن/ جلوگیری از B» دربرمی‌گیرد. در این رابطه، شروع A قبل از شروع B است، اما هیچ محدودیتی در پایان نسبی آن‌ها وجود ندارد که نشان دهد یک رابطهٔ علّی هر یک از دو رابطهٔ زمانی را دارد: قبل و همزمان.

BioCause یک چارچوب برچسب‌گذاری برای روابط علّی در متون زیست‌پزشکی ارائه می‌دهد و محدوده‌های استدلالی و جهت علّیت را مشخص می‌کند. در این رابطه، Dunietz, Levin and Carbonell (2017) تعمیم BioCause به حوزه‌های وسیع‌تر را پیشنهاد کردند. این روش چهار زیررابطهٔ علّی مختلف شامل انگیزه، استنتاج، هدف و نتیجه را متمایز می‌کند.

افزون بر مجموعهٔ انگلیسی، مجموعه‌های مشابهی برای سایر زبان‌ها نیز وجود دارد. Rehbein and Ruppenhofer (2020) منبع جدیدی برای روابط علّی آلمانی ارائه می‌کنند. در پژوهش آن‌ها از شمای برچسب‌گذاری Dunietz, Levin and Carbonell (2015) اقتباس شده است، اما با مجموعهٔ بسط‌یافته‌ای از آرگومان‌ها[1]. این مجموعه داده شامل ۴۳۹۰ مورد برچسب‌گذاری‌شده برای بیش از ۱۵۰ محرک مختلف است. شمای برچسب‌گذاری سه نوع رویداد علّی را که شامل نتیجه، انگیزه و هدف است، متمایز می‌کند. آن‌ها همچنین برچسب‌گذاری برای نقش‌های معنایی، یعنی علت و معلول برای رویداد علّی و بازیگر و طرف معلول را ارائه می‌دهند. این تحقیق از یک برچسب توالی علّی مبتنی بر «برت»[2] به‌عنوان معیار استفاده می‌کند.

افزون بر این مجموعه داده، پژوهش Sadek and Meziane (2018) مجموعه داده (SACB)[3] را معرفی می‌کند. این مجموعه، مجموعه جدیدی است که به روابط علّی عربی اختصاص دارد. این پیکره شامل مجموعه‌ای از جملات برچسب‌گذاری‌شده است که هر کدام نمونه‌ای از یک عنصر علّی دارند. این مجموعه با مثال‌هایی حاوی کلماتی که با پیشوندهای خاص به همراه استدلال‌های علت و معلولی پیشوند شده‌اند، برچسب‌گذاری شده است.

آمار این مجموعه در جدول ۱، نشان داده شده است. ردیف آخر این جدول PerCaus را نشان می‌دهد که در این مقاله معرفی شده و با مجموعه‌های موجود مقایسه شده است.

همچنین، برخی از مجموعه‌های خاص حوزهٔ علّی در حوزهٔ زیست‌پزشکی به بیماری‌ها، علائم و داروها اختصاص داده شده است؛ مانند کادِک[4] که در پژوهش Karimi et al. (2015) معرفی کرده است. این مجموعه متشکل از ۱۲۵۳ پست (۷۰۰۰ جمله) از انجمن پزشکی «آسکاپی‌شنت»[5] است که دارو، عوارض جانبی، بیماری، علائم و یافته‌های متنی

1. Arguments
2. BERT
3. Salford Arabic Causal Bank (SACB)
4. corpus of adverse drug event annotations (CADEC)
5. AskaPatient

را تعیین می‌کند. مجموعه‌های مشابهی که Leaman, Christopher and Graciela (2009); Gurulingappa et al. (2012); Deleger et al. (2012) در این زمینه ارائه کرده‌اند، به‌دلیل دامنه و برچسب‌های خاصی که دارند با موضوع این پژوهش چندان مرتبط نیستند و به ذکر اسامی آن‌ها بسنده کردیم.

مجموعه‌های دیگری که با علّیت مرتبط هستند، مجموعهٔ داده‌های استلزام متنی از جمله جفت‌های جمله با برچسب استلزام، تقابل و خنثی هستند. از آنجا که علّیت نوعی استلزام است، برخی از روابط استلزامی در این مجموعه داده روابط علّیت است. به‌عنوان مثال، جملات «حسن دیگر گرسنه نیست» و «حسن فقط ناهار خورد» دارای روابط استلزام و علّی است.

مجموعهٔ مسابقات شناسایی استلزام متنی[1] شامل دسته‌های آزمون و توسعه از سال ۲۰۰۵ تا ۲۰۱۰ است. این مسابقات از سال ۲۰۱۰ به‌عنوان کنفرانس تحلیل متن[2] و مسابقات semEval ادامه یافته است. در این راستا، مجموعه SICK[3] شامل ۱۰۰۰۰ جفت جملهٔ انگلیسی از دو منبع توضیحات ویدئو ImageFlickr و SemEval 2015 است. این منابع به‌صورت دستی با سه برچسب استلزام، تقابل و خنثی برچسب‌گذاری می‌شوند. سرانجام این‌که مجموعهٔ SNIL[4] دانشگاهٔ استنفورد شامل ۵۰۰۰۰۰ جفت جمله است که به‌صورت دستی در سه دسته برچسب‌گذاری شده است: استلزام، تقابل و خنثی.

برخی مسائل در مورد مجموعه دادگانی که در اغلب مطالعات فوق‌الذکر مورد بحث قرار گرفت، وجود دارد. اولین مسئله این است که مؤلفان یک عنصر علّی را پیوسته و بدون شکاف بین کلمات آن فرض می‌کنند، اما گاهی یک عنصر علّی (مثلاً علت) به دو یا چند قسمت (کلمات یا توکن‌ها) تقسیم می‌شود که در جمله همسایه و متوالی نیستند. نمونه‌ای از این مورد در جدول ۳، ارائه شده است. مسئلهٔ دوم این‌که، الگوهای در نظر گرفته‌شده، فقط به افعال علّی، نشانه‌های علّی یا الگوهای دستوری محدود، محدود می‌شوند. این موارد در مورد زبان‌هایی که ترتیب کلمات آزاد دارند، به مشکلاتی منجر می‌شود. برای نمونه، «سیگار کشیدن احتمال سرطان را افزایش می‌دهد» یک نمونهٔ علّی است که الگوهای علّی ساده نمی‌توانند آن را پوشش دهند.

1. recognizing textual entailment (RTE)
2. text analysis conference (TAC)
3. Sentences Involving Compositional Knowledge, available in: http://clic.cimec.unitn.it/composes/sick.html
4. http://nlp.stanford.edu/projects/snli/

در ایـن مقالـه، روش پیشـنهادی خـود را بـرای آماده‌سـازی یـک مجموعـه دادهٔ برچسب‌خوردهٔ علّیـت بـا دو ویژگـی مهـم مـورد بحـث قـرار می‌دهیـم. نخسـت این‌کـه هیـچ محدودیـت از پیـش تعریف‌شـده‌ای در الگوهـای دسـتوری وجـود نـدارد کـه نشـان دهـد هـر الگـوی علّـی می‌توانـد در PerCause رخ دهـد. ثانیـاً، عناصـر علّـی بیـش از یـک جـزء داشـته و اجـزا در جملـه می‌تواننـد فاصلـه داشـته باشـند؛ به‌طـوری کـه محدودیتـی در محـل وقـوع آن‌هـا در جملـه وجـود نـدارد. مجموعـه دادهٔ PerCause بـرای عناصـر علّـی برچسـب (درون–بیـرون–شـروع)[1] دارد (Ramshaw and Marcus 1995).

جدول ۱. مشخصات مجموعه‌های علّیت

پیکره	تعداد روابط علّی	مشخصات منبع داده	جزئیات برچسب‌گذاری
BECAUSE	۱۸۰۳	۵۳۸۰ جمله از NYTimes، Penn Treebank و دیگر مقالات خبری	نتیجه، انگیزه و انواع هدف برای روابط علّی
BioCause	۸۵۱	۱۹ متن کامل با دسترسی باز مجلات زیست‌پزشکی	علت رویداد، نوع، تم و علت برای هر رابطه
CaTeRS	۴۸۸	۳۲۰ داستان (۱۶۰۰ جمله) از مجموعهٔ ROCStories	چهار نوع از روابط علّی علت، فعال‌سازی، جلوگیری و علت برای پایان با فرض این‌که A سبب B است/ آن را فعال یا از آن جلوگیری می‌کند.
نسخهٔ ۱ BeCause	۴۰۰	۱۲۰۰ جمله از بخش واشنگتن از نیویورک تایمز	انواع انگیزه، استنتاج، هدف و نتیجه برای رابطه علّی
SemCore2.1[2]	۱۰۶۸	۳۵۲ متن از مجموعه Brown	برچسب‌های علّی یا غیرعلّی
پیکره علّی آلمانی	۴۳۹۰	۱۵۰ متن از دو منبع: (۱) متن روزنامه از مجموعه TiGer و (۲) سخنرانی‌های سیاسی از مجموعه Europarl	سه نوع رابطه علّی (نتیجه، انگیزه و هدف)
SACB	۲۱۶۲	۶۹۵۷۳ توکن از مجموعهٔ عربی (دسته روزنامه‌هایی که شامل تقریباً ۱۳۵ میلیون کلمه از مقالاتی که در سال‌های ۱۹۹۶ تا ۲۰۱۰ در کشورهای مختلف عربی منتشر شده است)	برچسب‌های علّی یا غیرعلّی
PerCause	۵۱۲۸	۴۴۴۶ جمله (۱۲۹۰۰۰ توکن) از مجموعهٔ Bijankhan و کتاب‌های عمومی	علت، معلول، و نشانه برای هر رابطهٔ علّی

1. inside–outside–beginning (tagging) (IOB)
2. http://lit.csci.unt.edu/~rada/downloads/semcor/semcor2.1.tar.gz

۳. روش پژوهش

۳-۱. پیکرهٔ برچسب‌گذاری‌شدهٔ علّیت PerCause

PerCause، پیکرهٔ برچسب‌گذاری‌شدهٔ علّیت توسعه‌یافته در این پژوهش شامل تقریباً ۱۲۹۰۰۰ توکن و ۵۱۲۸ رابطهٔ علّی است. مجموعهٔ داده‌های اولیه و مجموعهٔ برچسب‌های مورد استفاده در مجموعه زیربخش‌های زیر معرفی می‌شوند.

۳-۱-۱. منبع دادگان

مجموعهٔ دادگان خام اولیه برای برچسب‌گذاری از دو پیکرهٔ متفاوت انتخاب شده است: ۱) پیکرهٔ Peykareh و ۲) پیکرهٔ کتاب (که توسط محقق ایجاد شده است). پیکره Peykareh که در پژوهش (Bijankhan 2004) معرفی شده، تقریباً از ۱۰ میلیون توکن با برچسب اجزای کلام[1] تشکیل شده است (در این پژوهش فقط از دادگان خام استفاده شده است). اگرچه «پیکره» شامل متون ژانرهای مختلف است، اما غالب آن‌ها را مقالات خبری تشکیل می‌دهند. همچنین، از آنجا که مقالات خبری عناصر علّی عمومی را اغلب به‌درستی پوشش نمی‌دهند، ۱۰ رمان و کتاب عمومی فارسی را جمع‌آوری کرده و یک مجموعهٔ کلی شامل ۱/۸ میلیون توکن افزون بر مجموعه دادگان «پیکره» ساخته شد. «پیکرهٔ کتاب» برچسب برچسب اجزای کلام ندارد و برای این امر از یک برچسب‌گذار خودکار (توضیح در بخش ۴-۲) استفاده شد.

۳-۱-۲. مجموعهٔ برچسب‌ها

مجموعهٔ برچسب‌ها شامل سه برچسب «علت»، «معلول» و «نشانهٔ علّی» است. با توجه به شباهت برچسب‌گذاری علّیت به وظیفهٔ شناسایی موجودیت‌های نامدار[2] یا قطعه‌بندی[3]، فرمت IOB برای این کار انتخاب شد. جدول ۲، تعاریف و مثال‌هایی را برای این سه برچسب نشان می‌دهد. برخی از نمونه‌های برچسب‌گذاری در جدول ۳، و شکل ۱، نشان داده شده است.

1. part of speech (POS) 2. named entity recognition 3. chunking

جدول ۲. تعریف برچسب‌ها و مثال

برچسب	تعریف	مثال
علت	یک شخص، رویداد، عمل یا چیزی که باعث ایجاد عمل، پدیده یا شرایطی می‌شود.	سیگار کشیدن ریسک سرطان را بالا می‌برد.
معلول	تغییری که نتیجه یا عاقبت یک عمل یا علت است.	سیگار کشیدن ریسک سرطان را بالا می‌برد.
نشانهٔ علّی	کلمه یا عبارتی که رابطه علّی را مشخص می‌کند.	سیگار کشیدن ریسک سرطان را بالا می‌برد.

جدول ۳. نمونه‌ای از دادهٔ برچسب‌خورده

جملهٔ بدون برچسب	کمبود هورمون کورتیزول باعث خشونت‌طلبی در برخی از پسران می‌شود؛ خصوصاً در سنین نوجوانی.
جملهٔ برچسب‌خورده	<entity="cause-1">کمبود هورمون کورتیزول</entity> <entity="Causal-mark-1">باعث</entity> <entity="effect-1">خشونت‌طلبی در برخی از پسران</entity> می‌شود؛ <entity = "effect t-1">خصوصاً در سنین نوجوانی</entity>.

```
۱  کمبود هورمون کورتیزول باعث خشونت طلبی در
۲  برخی از پسران می شود ، خصوصا در سنین نوجوانی .
۳
۴  کمبود ⟵ B-Cause
۵  هورمون ⟵ I-Cause
۶  کورتیزول ⟵ I-Cause
۷  باعث ⟵ B-Mark
۸  خشونت طلبی ⟵ B-Effect
۹  در ⟵ I-Effect
۱۰  برخی ⟵ I-Effect
۱۱  از ⟵ I-Effect
۱۲  پسران ⟵ I-Effect
۱۳  می شود ⟵ O
۱۴  ، ⟵ O
۱۵  خصوصا ⟵ I-Effect
۱۶  در ⟵ I-Effect
۱۷  سنین ⟵ I-Effect
۱۸  نوجوانی ⟵ I-Effect
۱۹  . ⟵ O
```

شکل ۱. نمونه‌ای از فرمت IOB

۳-۱-۳. آماده‌سازی نیمه‌خودکار پیکرهٔ اولیهٔ PerCause

پس از تهیهٔ مجموعهٔ خام اولیه از «پیکره» و تجمیع آن با مجموعهٔ دادگان کتاب، زیرمجموعه‌ای از این مجموعه داده برای برچسب‌گذاری دستی به‌صورت خودکار انتخاب می‌شود. در این بخش فرایندهای انجام‌شده برای ساخت این زیرمجموعه (یعنی پیکرهٔ خام) بیان می‌شود. بعد از انجام این مراحل، پیکره برای برچسب‌گذاری آماده است.

◆ **نرمال‌سازی**

از آنجا که برخی از کاراکترها در صفحه‌کلید فارسی بیش از یک یونی‌کد متناظر دارند (مانند «ی» و «ک»)، کاراکترها در این مرحله یکسان می‌شوند.

◆ **جداسازی جملات**

متن، پس از نرمال‌سازی به جملات تقسیم می‌شود. ما از علائم نگارشی مانند «.»، «!» و «؟» به‌عنوان جداکننده استفاده می‌کنیم. در ضمن، قوانینی در استفاده از این نمادها در کلمات خاص (مانند کلمات اختصاری)، اعداد و «یوآرال»[1]ها در نظر گرفته شده تا از تقسیم‌بندی نادرست جلوگیری شود.

◆ **توکن‌بندی**

فرایند توکن‌بندی برای تبدیل هر جمله به دنباله‌ای از توکن‌ها مانند کلمات، علائم نگارشی و اعداد انجام می‌شود. نشانه‌گذاری در زبان فارسی در مقایسه با زبان انگلیسی از آنجا که فاصله در زبان فارسی یک جداکنندهٔ قطعی نیست و ممکن است در یک کلمه رخ دهد و یا دو کلمهٔ مجزا بدون هیچ فاصله‌ای در بینشان ظاهر شوند، چالش‌برانگیزتر است. در این مرحله، از جعبه‌ابزار STeP-1 برای توکن‌بندی جملات استفاده شده است که توسط Shamsfard, Jafari and Ilbeygi (2010) ایجاد شده است.

◆ **فیلتر کردن طول**

پس از تقسیم متن به جملات و جملات به توکن، جملات بسیار طولانی و بسیار کوتاه را حذف می‌کنیم و جملاتی با طول بین ۵ تا ۱۰۰ نشانه را نگه می‌داریم. مشاهدات ما نشان می‌دهد که جملاتی با کمتر از پنج یا بیش از ۱۰۰ نشانه یا حاوی هیچ رابطهٔ علّی نیستند یا احتمالاً با ادغام یا تقسیم نادرست بخش‌ها ساخته شده‌اند، و بنابراین می‌شود آن‌ها را نادیده گرفت. در پایان این بخش میزان اتلاف جمله مورد بحث قرار گرفته است.

◆ **انتخاب جملات کاندید**

در اینجا جملاتی که با الگوهای از پیش تعریف‌شدهٔ علّیت مطابقت دارند، نامزد علّی نامیده می‌شوند. مجموعهٔ نهایی برای برچسب زدن دستی را از میان این مجموعه کاندید

1. URL

انتخاب می‌کنیم. برای این کار فهرستی از ۳۰ الگوی علّی (یا عباراتی را که به‌نحوی نشانهٔ رابطهٔ علّی هستند)، تهیه کردیم که برخی از آن‌ها در جدول ۴، نشان داده شده است. در جدول ۴، نماد #به این معناست که کلمات الگو باید به‌ترتیب تعریف‌شده دیده شوند، اما لزوماً یک عبارت ایجاد نمی‌کنند.

جدول ٤. نمونه‌ای از الگوهای علّی

دسته	الگو	درصد (به نسبت کل الگوها)
فعل	(تولید کردن، افزایش دادن، منجر شدن، بالا بردن، موجب#شدن، تغییر دادن)	۲۳
حرف اضافه	(به‌وسیله، اگر#آنگاه، چون، در نتیجه، زیرا، از آنجا که)	۵
اسم	(حاصل، سبب، باعث، موجب)	۴۵
ترکیبی	(علت#درصد، به این بابت، به‌علت، به‌دلیل، حاصل از، حاصل#این#آن، علت#درصد، از#برای#استفاده، ناشی از، به‌همین علت، به همین دلیل، احتمال#افزایش، عوامل#مهم، به این دلیل)	۲۷

برای ایجاد این لیست در وهلهٔ اول دو لیست را به‌صورت دستی آماده می‌کنیم: ۱) الگوهای اولیه (حدود ۱۰ الگوی علّی)، و ۲) جفت‌های علّی اولیه (حدود ۸۰ جفت کلمه یا عبارت علت و معلول). سپس، جفت‌های علّی اولیه را در مجموعهٔ جملات از پیش پردازش‌شده جست‌وجو کرده و جملاتی را که شامل این جفت‌های علّی است، استخراج می‌کنیم. از طرف دیگر، الگوهای اولیه را در جملات جست‌وجو کرده و از جملات استخراج‌شده جفت‌های علّی بیشتری استخراج می‌کنیم. سپس، مرحلهٔ اول را با جفت‌های علّی استخراج‌شدهٔ جدید تکرار می‌کنیم و تکرار را ادامه می‌دهیم. فهرست الگوهای علّی از بررسی ساختار جملاتی به‌دست آمده که الگوهای علّی را داشته و سپس، یک‌به‌یک به‌صورت دستی بررسی شده‌اند. مجموعهٔ جملاتی که با الگوهای علّی مطابقت دارند، به‌عنوان نامزدهای علّی اولیه انتخاب می‌شوند. این روش در شکل ۲، بیان شده است.

- لیست اولیهٔ الگوها و لیست جفت عبارت علّی را به‌صورت دستی ایجاد کنید.
- تا وقتی جفت علّی جدید یا الگوی علّی پیدا شود:
  - الف. در میان جملات، لیست جفت عبارات علّی را جست‌وجو و الگوهای علّی بیشتری استخراج کنید.
  - ب. لیست الگوهای علّی را در جملات جست‌وجو و جفت‌های علّی جدید را پیدا کنید.
  - پ. هر دو لیست را به‌روزرسانی کنید.
- مجموعه‌ای از جملاتی را که با الگوهای علّی تطبیق دارند، به‌عنوان کاندید علّی مناسب برای برچسب زدن انتخاب کنید.

شکل ۲. شبه‌کد انتخاب جمله‌های علّی مناسب برای برچسب‌گذاری

◆ **انتخاب مجموعهٔ جملات نهایی برای برچسب‌گذاری**

پس از آماده کردن مجموعهٔ موارد علّی اولیهٔ مناسب، جفت‌های علّی استخراج‌شده بازنگری و جفت‌های نامناسب حذف می‌شوند. برای مثال، برخی از جفت‌های علت و معلولی دائمی نیستند و فقط در یک جملهٔ خاص علت و معلول هستند؛ مثلاً «به‌دلیل بیماری به پاریس رفت». این نوع جملات در مقالات خبری فراوان دیده می‌شود. از آنجا که می‌خواهیم جملات کلی باشند و در همهٔ زمینه‌ها دقت بالایی داشته باشند، بیشتر در جست‌وجوی مثال‌هایی هستیم که علت دلیل مستقیم معلول در جمله باشد. بنابراین، نمونه‌ها را در مجموعه بررسی می‌کنیم، برخی از الگوها را ساده می‌کنیم، و مجموعهٔ نهایی را برای برچسب‌گذاری انتخاب می‌کنیم.

در طول سه مرحلهٔ اول پیش‌پردازش (یعنی نرمال‌سازی، توکن‌بندی و تقسیم‌بندی به جمله‌ها) هیچ جمله‌ای از دست نمی‌رود؛ هرچند در مرحلهٔ فیلتر طول حدود ۱۰ درصد اتلاف داریم. پس از انتخاب موارد علّی مناسب (۶ درصد جملات ابتدایی دارای الگوهای علّی هستند)، حدود ۵۰ درصد از جملات را به‌عنوان مورد مناسب علّی که دارای روابط علّی هستند، انتخاب می‌کنیم. جدول ۵، مشخصات مجموعهٔ اولیه را قبل از برچسب‌گذاری علّی نشان می‌دهد.

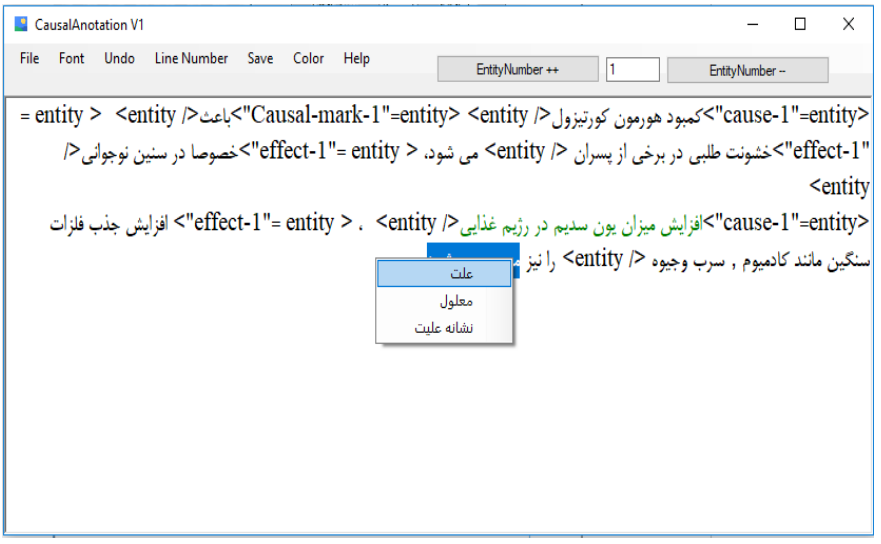
جدول ۵. مشخصات مجموعهٔ اولیه قبل از برچسب‌گذاری

	تعداد جملات	تعداد کلمه‌ها
پیکرهٔ کاندید اولیه برای علّیت	۴۴۴۶	۱۲۹۲۹۳

۳-۱-٤. برچسب‌گذاری پیکره

پس از این مراحل، جملات این مجموعه آمادهٔ برچسب‌گذاری هستند.

برای این کار، یک ابزار برچسب‌گذاری ایجاد کردیم تا به زبان‌شناسان کمک کنیم که جملات را به‌صورت دستی برچسب‌گذاری کنند. این ابزار به برچسب‌گذار اجازه می‌دهد که یک عنصر چندکلمه‌ای را (حتی با وجود فاصلهٔ بین آن‌ها) برچسب‌گذاری کند. همچنین، چندین برچسب گذار می‌توانند به‌طور همزمان از آن استفاده کنند. شمایی از رابط کاربری گرافیکی آن در شکل ۳، نشان داده شده است. کافی است برچسب‌گذارها فقط روی یک عبارت یا کلمه کلیک راست کرده و برچسب مربوطه را انتخاب کنند. افزون بر این، برچسب‌گذار می‌تواند یک عنصر چندبخشی را با استفاده از گزینهٔ entity number برچسب‌گذاری کند. پس از برچسب‌گذاری داده‌ها برای استفادهٔ بیشتر به فرمت IOB در ساختار داخلی تبدیل می‌شوند.



شکل ۳. نمایی از رابط کاربری گرافیکی برچسب‌گذاری

جدول ۶، مشخصات برچسب‌های علّی B-Cause، I-Cause، B-Effect، I-Effect، B-Mark و I-Mark را در پیکره نشان می‌دهد. از مجموع ۱۲۹۰۰۰ توکن در پیکره، حدود ۶۳۰۰۰ توکن علّی (توکن‌هایی با برچسب‌های علّی) در قالب ۵۱۲۸ رابطهٔ علّی (سه‌گانه یا چندتایی علّی) در ۴۴۴۶ جمله داریم. قابل توجه است که برخی از جملاتِ کاندید، هیچ رابطهٔ علّی ندارند.

جدول ٦. اطلاعات آماری برچسب‌ها در پیکره

	برچسب	تعداد رخداد	درصد نسبت به کل توکن‌ها در پیکره
۱	B-cause	۴۲۶۱	
	I-cause	۱۹۶۸۳	
	تعداد کل علت‌ها	۲۳۹۴۴	۱۷/۹۶
۲	B-Effect	۴۵۹۷	
	I-Effect	۲۷۱۸۰	
	تعداد کل معلول‌ها	۳۱۷۷۷	۲۳/۸۴
۳	B-Causal mark	۵۱۲۸	
	I-Causal mark	۲۱۵۷	
	تعداد کل نشانه‌های علّیت	۷۲۸۵	۵/۴
۴	تعداد کل	۶۳۰۰۶	۴۷/۲

۳-۱-۵. توافق برچسب‌گذاری

توافق برچسب‌گذاری یک فاکتور اساسی در توسعهٔ یک پیکره است. اختلاف بین برچسب‌گذارهای مختلف در برچسب‌گذاری دستی یک پیکره اجتناب‌ناپذیر است و حتی مشاهده می‌شود که یک برچسب‌گذار یک متن واحد را در دو زمان مختلف به‌طور متفاوت برچسب‌گذاری می‌کند. بنابراین، برای اطمینان از صحت نتایج برچسب‌گذاری باید توافق بین برچسب‌گذاران محاسبه شود. افزون بر این، تجربه نشان داده است که استفاده از مجموعهٔ یکنواخت‌تر و دقیق‌تر در سیستم‌های مبتنی‌بر یادگیری ماشین به نتایج بهتری منتهی می‌شود. در این راستا معیارهای مختلفی برای سنجش میزان توافق برچسب‌گذار استفاده می‌شود. یکی از معیارهای معروف و پرکاربرد برای این موضوع، «کاپا»[1] است که در پژوهش (Cohen 1960) مطرح شده است.

1. Kappa

این معیار یک معیار آماری است که در زبان‌شناسی پیکره‌ای برای بررسی میزان توافق بین دو برچسب‌گذار یک پیکره استفاده می‌شود. این معیار سعی می‌کند اثر برچسب‌هایی را که به‌طور تصادفی شبیه یکدیگر هستند، کاهش دهد. برای این منظور «کاپا» به صورت زیر محاسبه می‌شود:

$$k = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

در این فرمول، Pr(a) سطح درست شباهت بین دو برچسب‌گذار و Pr(e) میزان تشابه تصادفی بین دو برچسب‌گذار است. برای استفاده از معیار «کاپا»، در مواردی که بیش از دو برچسب‌گذار دارند، این مقدار را برای هر جفت برچسب‌گذار محاسبه می‌کنیم و میانگین مقادیر آن‌ها را می‌گیریم. در زبان‌شناسی ضرایب «کاپا» بالای ۸۰ درصد برای یک مجموعهٔ مناسب در نظر گرفته می‌شود (Green 1997).

در اینجا، برای کاهش ابهامات در روش برچسب‌گذاری، یک دستورالعمل برچسب‌گذاری دقیق تهیه شد و برای این فرایند از دو برچسب‌گذار استفاده شد. برای این منظور، ۱۳۰۰ کلمه از پیکره به‌صورت تصادفی انتخاب شده و دوبار برچسب‌گذاری گردید. توافق برچسب‌گذاری ۹۴/۵ درصد محاسبه شده است، به‌طوری که برای مجموعهٔ برچسب‌گذاری قابل قبول است و اکنون پیکره برای استفادهٔ سیستم‌های شناسایی علّیت آماده است.

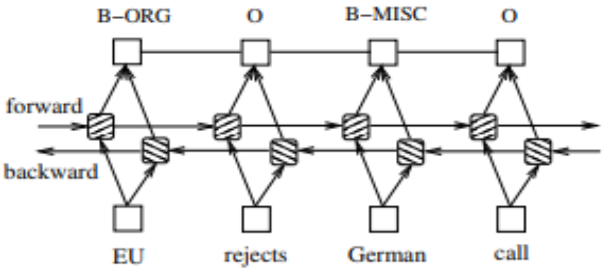
۳-۲. سیستم شناسایی مرز عناصر علّی

هدف این پژوهش ایجاد یک پیکرهٔ فارسی دارای برچسب علّی و معرفی بستر آزمونی برای تشخیص زوج‌های علّی و مرزهای آن‌ها با استفاده از پیکرهٔ ساخته‌شده است. به این منظور از پیکرهٔ علّی ایجادشده برای آموزش یک سیستم پایهٔ یادگیری ماشین و دو سیستم یادگیری عمیق استفاده شده و بستر آزمونی در این حوزه برای دیگر پژوهشگران ایجاد شده است. در این بخش روش‌های طبقه‌بندی و ویژگی‌های مورد استفاده برای ساخت بستر آزمون مورد بحث مطرح می‌شود. به‌عبارت دیگر، مسئلهٔ شناسایی مرز عناصر علّی به‌عنوان یک مسئلهٔ طبقه‌بندی چندکلاسه مطرح می‌شود که کلاس‌های آن برچسب‌های عناصر علّی هستند. بنابراین، این بخش قسمت‌های طبقه‌بندی و ویژگی‌های مورد استفاده در سیستم را توضیح می‌دهد.

۳-۲-۱. روش‌های طبقه‌بندی

ما بخشی از پیکره را به‌عنوان دادهٔ آموزشی جدا کردیم و چند طبقه‌بندی‌کننده مانند Naïve Bayes، RBF و CRF را برای توسعهٔ بستر آزمون[1] آزمایش کردیم. همچنین، دو سیستم یادگیری عمیق با بازنمایی‌های مختلف پیاده‌سازی شده است. این سیستم‌ها با استفاده از مجموعهٔ داده‌های مشابه، آموزش داده شده و آزمایش می‌شوند.

۱. اولین سیستم (شکل ۴) یک چارچوب یادگیری عمیق (bi-LSTM+CRF) با استفاده از Tensorflow است که در مقالهٔ Huang, Wei and Kai (2015) معرفی شده است[2].



شکل ٤. مدل BI-LSTM-CRF در مقالهٔ (Huang, Wei and Kai, 2015)

روال این سیستم به‌شرح زیر است:

- ◇ حالت‌های نهایی یک bi-LSTM در جاسازی کاراکترها به هم وصل می‌شود تا نمایشی مبتنی بر کاراکتر از هر کلمه به‌دست آید.
- ◇ این بازنمایی به یک نمایش برداری استاندارد کلمه الحاق می‌شود.
- ◇ روی هر جمله یک bi-LSTM اجرا می‌شود تا یک بازنمایی متنی از هر کلمه استخراج شود.
- ◇ با CRF زنجیرهٔ خطی رمزگشایی انجام می‌شود.

در این مرحله برای جاسازی کلمه از روش «فستتکست» و مدل سی‌بو[3] در مجموعهٔ «ویکی‌پدیا»ی فارسی با بیش از یک میلیارد توکن استفاده می‌شود.

۲. با توجه به ظهور مدل‌های «برت» (که نخستین‌بار در مقالهٔ Delvin et al. (2018) مطرح شدند) و عملکرد مثبت آن‌ها در کارهای مشابه، یک مدل از پیش آموزش‌دیده را نیز روی داده‌های تولیدشده تنظیم می‌کنیم.

1. benchmark
2. https://github.com/guillaumegenthial/sequence_tagging
3. CBOW: Continious bag of words

نتایج و جزئیات پیاده‌سازی‌ها در مورد طبقه‌بندی‌کننده‌های عمومی و سیستم‌های عمیق در بخش ۲-۵ ارائه شده است.

۳-۲-۲. ویژگی‌ها

مدل تشخیص مرز عناصر علّی با استفاده از ویژگی‌های اصلی کلمه، ریشهٔ کلمه و برچسب‌های POS آموزش داده شد. در این پژوهش از ریشه‌یابی در درجهٔ اول برای پوشش صرف افعال مختلف در جمله‌ها استفاده می‌شود. همچنین، گاهی برچسب‌های POS خاص در مجاورت علامت علّی نشان‌دهندهٔ یک الگوی رابطهٔ علّی است. بنابراین، ویژگی POS نیز مورد استفاده قرار گرفته است. برچسب‌های POS در مرحلهٔ آموزش به‌عنوان ویژگی در کنار دادگان و برچسب‌ها به طبقه‌بندی‌کننده‌ها وارد می‌شوند.

افزون بر این، دو مدل برچسب‌گذاری POS با مجموعهٔ ۱۶ برچسب و ۱۰۰ برچسب بر روی مجموعهٔ «پیکره» به‌ترتیب با دقت ۹۴ و ۹۰ درصد آموزش داده شدند. در ضمن، یک الگوریتم سبک ریشه‌یابی برای سیستم پایه اعمال شد که به‌سادگی پیشوندها و پسوندها را حذف می‌کند. تأثیر در نظر گرفتن برچسب‌های POS و ریشه‌یابی در بخش ۴-۴ ارزیابی شده است.

٤. تجزیه و تحلیل یافته‌ها

٤-١. معرفی داده‌ها و معیارهای ارزیابی

برای پوشش اثر اندازهٔ دادگان در تحلیل، آزمایش‌ها را روی دو مجموعهٔ داده با در نظر گرفتن میزان داده‌های برچسب‌گذاری‌شدهٔ موجود در ترتیب زمانی انجام دادیم: ۱) مجموعهٔ دادهٔ اول با ۴۷۰۰۰ توکن، و ۲) مجموعهٔ دادهٔ دوم با ۱۲۹۰۰۰ نشانه (شامل جملات مجموعهٔ اول)

برای مجموعهٔ اول، مجموعهٔ دادهٔ برچسب‌گذاری‌شدهٔ خود را به سه دسته (آموزش، آزمایش و توسعه) تقسیم کردیم که هر کدام به‌ترتیب، ۸۰، ۱۰ و ۱۰ درصد از داده‌ها را شامل می‌شوند. اما، برای مجموعهٔ دوم از قسمت غیرهمپوشان، فقط یک مجموعه توسعه جدا شد و بقیه به مجموعهٔ دادهٔ آموزشی اضافه شد. از مجموعهٔ دادهٔ قبلی که شامل ۴۷۵۰ توکن است، برای آزمایش استفاده شد. افزون بر این، از یک روش اعتبارسنجی متقاطع ۱۰ مرحله‌ای[1] برای اطمینان از دقت نتایج استفاده شد.

1. 10-fold cross validation

معیارهـای اسـتاندارد صحـت (Acc)، دقـت (P)، فراخوانـی(R) و F-Measure (F) بـرای ارزیابـی عملکـرد سیسـتم اسـتفاده شـده اسـت. ایـن معیارهـا به‌صـورت زیـر تعریـف می‌شـوند:

$$\text{Accuracy (Acc)} = \frac{\text{all correct system decisions for a specific tag}}{\text{all samples}}$$

$$\text{Precision (P)} = \frac{\text{correct system decisions for a specific tag}}{\text{all system decisions}}$$

$$\text{Recall (R)} = \frac{\text{correct system decisions for a specific tag}}{\text{what system should have decided}}$$

$$F - measure = \frac{2 \cdot P \cdot R}{(P + R)}$$

ایـن معیارهـا بـرای هـر برچسـب محاسـبه می‌شـود و مقـدار میانگیـن هـر ۷ کلاس B-Cause، I-Couse، B-Effect، I-effect، B-Mark، I-mark و O بـرای مقایسـهٔ کلـی سیسـتم‌ها در نظـر گرفتـه می‌شـود.

٤-٢. ارزیابـی روش‌هـای مختلـف یادگیـری ماشـین بـر روی مجموعـهٔ پیشـنهادی علّیـت (PerCause)

بـرای طبقه‌بندی‌کننده‌هـای عمومـی یعنـی RBF و Naïve Bayes، از جعبه‌ابـزار WEKA[1] بـا پارامترهـای پیش‌فـرض آن اسـتفاده شـد و سـپس، توکن‌هـا بـه بردارهـای کلمـه تبدیـل شـد. بـرای پیاده‌سـازی طبقه‌بنـد CRF از PocketCRF اسـتفاده شـد و f (حـد آسـتانهٔ تکـرار) مطابـق بـا پیش‌فـرض توسـعهٔ طبقه‌بندی‌کننـده CRF روی ۴ قـرار داده شـد. همچنیـن، p تعـداد نخ‌هـا[2]ی پردازشـی را تعییـن می‌کنـد و بـرای بهبـود عملکـرد، m روی ۱ و p برابـر ۴ قـرار داده شـد کـه در اینجـا، وقتـی m برابـر ۱ اسـت، حافظـهٔ کمتـری اسـتفاده می‌شـود. افـزون بـر ایـن، اندازه‌هـای مختلـف پنجـره در ایـن فراینـد آزمایـش شـد.

در اولیـن پیاده‌سـازی یادگیـری عمیـق (bi-LSTM+CRF) ۴۵ دوره[3]، ۱۰۰ بعـد بـرای جاسـازی کاراکتـر و ۳۰۰ بعـد بـرای بـردار کلمـات اسـتفاده شـد.

بـرای دومیـن سیسـتم عمیـق (بـر پایـهٔ بـرت)، از «پارس‌بـرت»[4] کـه در پژوهـش Farahani et al. (2020) معرفـی شـده، اسـتفاده و مـدل بـا داده‌هـای فارسـی علّیـت تنظیـم شـد. بـرای تنظیـم

1. https://www.cs.waikato.ac.nz/ml/weka/
2. thread
3. epoch
4. ParsBert

دقیق[1] کردن، روش پیشنهادی Weizhepei (2020) مورد استفاده قرار گرفت. این روش یک راه‌حل مبتنی بر PyTorch برای وظیفهٔ شناسایی موجودیت‌های نامدار با مدل «برت» گوگل است. در پژوهش Weizhepei (2020)، نتیجهٔ F-measure در وظیفهٔ NER روی پیکرهٔ MSRA چینی ۹۴/۶۴ و روی مجموعهٔ دادهٔ انگلیسی Conll (2003) ۹۶/۴ است.

جدول‌های ۷ و ۸ ارزیابی عملکرد روش‌های مختلف یادگیری ماشین آموزش دیده با استفاده از مجموعه‌های دادهٔ اول و دوم را ارائه می‌دهند. جدول ۹، بهترین نتایج سیستم مرتبط با مجموعهٔ داده‌های آموزشی مختلف را نشان می‌دهد.

جدول ۷. مقایسهٔ عملکرد سیستم‌های مبتنی بر یادگیری ماشین روی مجموع دادهٔ ۱ (۴۷۰۰۰ توکن)

سیستم	دقت	فراخوانی	معیار F	صحت
CRF	۷۷	۷۵	۷۵/۵	۷۴/۷
RBF Network	۴۶/۴	۵۰	۴۸	۴۹/۹
Naïve Bayes	۴۶/۶	۴۸	۴۷/۲	۴۷/۹
یادگیری عمیق (bi-LSTM+CRF)	۶۵	۵۲	۵۸	۸۶
یادگیری عمیق (پارس‌برت)	۶۴	۷۱	۶۷	۸۸

جدول ۸. مقایسهٔ عملکرد سیستم‌های مبتنی بر یادگیری ماشین روی مجموعهٔ دادهٔ ۲ (۱۲۹۰۰۰ توکن)

سیستم	دقت	فراخوانی	معیار F	صحت
CRF	۷۹/۹	۷۲	۷۶/۱	۷۵/۵
RBF Network	۴۹/۳	۴۷/۹	۴۶/۲	۴۷/۸
Naïve Bayes	۵۱/۷	۵۰/۶	۵۰/۳	۵۰/۶
یادگیری عمیق (bi-LSTM+CRF)	۷۰/۹	۶۵	۷۱/۳	۹۱/۴
یادگیری عمیق (پارس‌برت)	۶۶	۷۶	۷۰/۶	۸۹

طبق جداول ۷ و ۸، اگر عملکرد کلی را در نظر بگیریم، سیستم CRF با معیار F برابر با ۷۶ از سایر سیستم‌های پیاده‌سازی شده بهتر عمل می‌کند. این مقدار با عدد ۸۰ گزارش شدهٔ Girju در پژوهش سال ۲۰۰۳، عدد ۸۱ Chang و Choi در ۲۰۰۵ و سیستم عربی Sadek

1. fine tune

and Meziane (2018) با بهترین نتیجهٔ ۷۶ روی داده‌های تست خودشان قابل مقایسه است. جزئیات ارزیابی CRF در بخش بعدی ارائه شده است.

Rehbein and Ruppenhofer (2020) در تعیین برچسب‌های علّی در مجموعه آلمانی خود برای سیستم‌های عمیق به معیار F برابر با ۷۲/۲ رسیدند. این نتیجه با نتیجه‌ای که در سیستم یادگیری عمیق ما به‌دست آمده، کاملاً قابل مقایسه است. اگرچه در مقالهٔ Weizhepei (2020)، جایی که پیاده‌سازی مبتنی بر «برت» بر روی یک وظیفهٔ مشابه (NER) اعمال شده، نتیجهٔ معیار F برابر با ۹۴/۶ روی پیکرهٔ MSRA چینی و ۹۶/۴ در مجموعهٔ دادهٔ انگلیسی Conll (2003) بوده، این نتایج احتمالاً به‌دلیل بزرگ و غنی بودن مجموعهٔ دادگان رخ داده است. افزون بر این، این بالا بودن نتایج ممکن است به‌دلیل نبودِ وابستگی‌های دور و موجودیت‌های چندقسمتی نیز باشد.

نکتهٔ قابل توجه در مورد جدول‌های بالا این است که افزایش اندازهٔ داده‌ها هیچ تغییر قابل ملاحظه‌ای در نتایج CRF ندارد، اما دقت و معیار F را برای سیستم یادگیری عمیق (bi-LSTM+CRF) با دادگان آموزشی مشابه تا ۱۳ درصد بهبود می‌بخشد که پیشرفت قابل توجهی است.

در سیستم مبتنی بر «برت» نتایج نهایی نزدیک به دیگر سیستم‌های یادگیری عمیق است و عملکرد سیستم با مجموعهٔ دادهٔ اول و دوم تقریباً مشابه است. این نتیجه تأیید می‌کند که «برت» با داده‌های کمتر به‌خوبی کار می‌کند و میزان افزایش داده در مجموعهٔ دوم برای بهبود عملکرد سیستم «برت» کافی نبوده است. افزون بر این، وجود لایهٔ CRF در سایر شبکه‌های آموزش‌دیدهٔ عمیق، آن را در معیار F، حداقل در این حجم از داده‌های آموزشی، اندکی برتری داده است.

٤-٣. ارزیابی اثر افزایش حجم دادگان به‌صورت خودکار

در بخش قبل مشاهده کردیم که اگرچه یادگیری عمیق (bi-LSTM+CRF) بهترین صحت را در بین سایر سیستم‌ها دارد، عملکرد آن بر اساس معیار F و دقت کمتر از CRF است. چون روش‌های یادگیری عمیق به داده‌های مقیاس بزرگ نیاز دارند، تصمیم بر این شد که با استفاده از دقیق‌ترین سیستم پیاده‌شده یعنی معماری bi-LSTM+CRF به‌طور خودکار داده‌ها را برای ارزیابی تأثیر افزایش اندازهٔ مجموعه به‌صورت خودکار تولید کنیم. برای این منظور، bi-LSTM+CRF برای تولید داده‌های خودکار انتخاب شد، زیرا

دقت در این بخش مهم‌تر از فراخوانی[1] است. بنابراین، داده‌های بیشتری به‌طور خودکار با این سیستم یادگیری عمیق برچسب‌گذاری شدند. خروجی سیستم bi-LSTM+CRF به‌صورت یک مجموعهٔ داده متشکل از ۴۰۰۰۰۰ توکن با دقت ۹۱ (به نام مجموعهٔ دادهٔ خودکار در جدول‌ها آورده شده) استفاده شد. مجموعهٔ داده‌های جدید تولیدشده به‌طور خودکار برای آموزش یک سیستم یادگیری عمیق استفاده شد. داده‌های مورد استفاده برای برچسب‌گذاری خودکار از میان مجموعه‌های «پیکره» و یک مجموعه دادهٔ خبری انتخاب شدند. افزون بر این، همهٔ جملات از دامنهٔ مشابه PerCause هستند.

جدول ۹، نتایج به‌دست‌آمده در مجموعه داده‌های مختلف را نشان می‌دهد. همان‌طور که مشاهده می‌شود، در سه ردیف آخر جدول از مجموعهٔ تولید خودکار برای آموزش سیستم با دقت ۹۱ درصد استفاده شده است.

جدول ۹. مقایسهٔ عملکرد سیستم‌های مبتنی بر یادگیری عمیق روی مجموعه دادگان متفاوت

سیستم	صحت	معیار F
مجموعه دادهٔ ۱	۸۶	۵۸/۷
مجموعه دادهٔ ۲	۹۱/۴	۷۱/۳
مجموعه دادهٔ تولید خودکار	۸۴/۴۷	۵۶/۱
مجموعه دادهٔ تولید خودکار + مجموعهٔ ۱	۸۰/۳۷	۵۷/۶
مجموعه دادهٔ تولید خودکار + مجموعهٔ ۲	۸۹/۷	۶۷

همان‌طور که جدول ۹، نشان می‌دهد، افزودن داده‌های تولیدشده به‌طور خودکار عملکرد سیستم را بهبود نمی‌بخشد. توضیح این است که این مجموعه دادهٔ تولیدشده به اندازهٔ کافی دقیق نیست که بتوان از آن به‌عنوان دادهٔ آموزشی استفاده کرد. در کارهای آینده می‌توان برای استفاده از این پایگاه دادهٔ تولید خودکار، روش‌هایی با نظارت ضعیف در نظر گرفت.

٤-٤. ارزیابی بهترین سیستم

همان‌طور که در بخش قبل نشان داده شد، با اندازهٔ دادگان فعلی، CRF بهترین عملکرد کلی را در بین تمام سیستم‌های آزمایش‌شده دارد؛ اگرچه معماری‌های یادگیری

1. recall

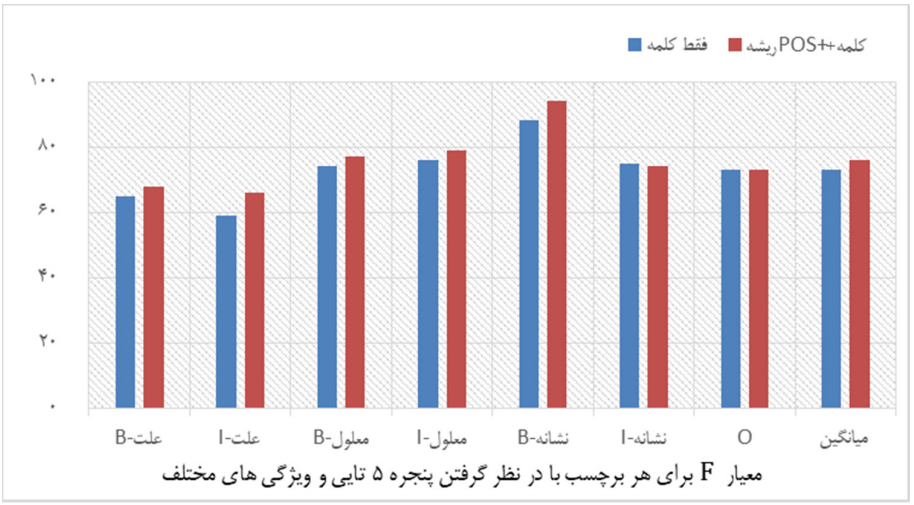
عمیق صحت بهتری نسبت به روش‌های دیگر دارد. این نتیجه تاحدی مورد انتظار است، زیرا CRF معیار F بهتری را در مقایسه با سایر روش‌های یادگیری ماشین در چنین مسائلی نشان داده است.

در این بخش، تأثیر ویژگی‌های مختلف بر عملکرد سیستم پایهٔ شناسایی علّیت با طبقه‌بند CRF برای هر کلاس بررسی می‌شود. در جدول ۱۰، نتایج سیستم CRF بدون هیچ عنصر افزوده (فقط کلمات و برچسب‌ها) و با دو ویژگی ساده برای هر کلاس ارائه شده است.

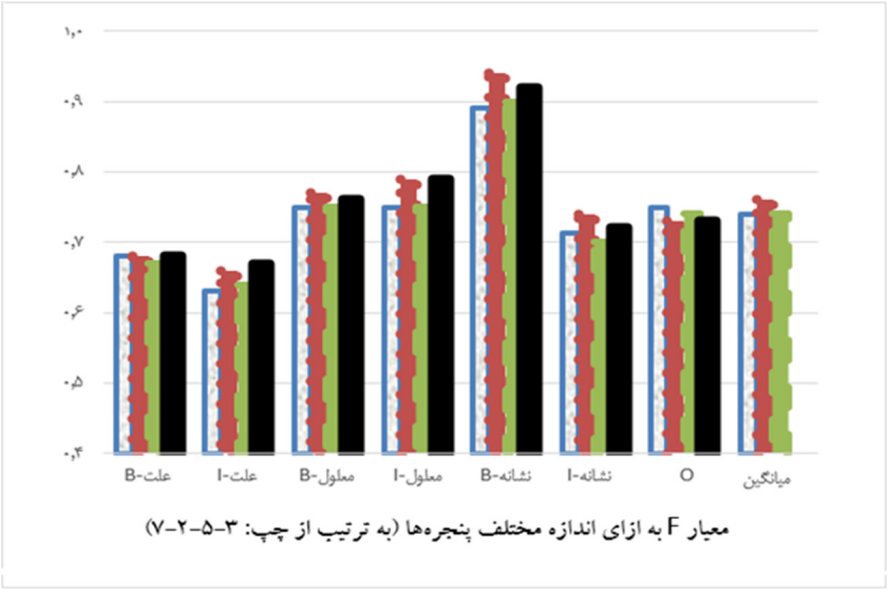
جدول ۱۰. مقایسهٔ عملکرد سیستم CRF با ویژگی‌های متفاوت

ویژگی	فقط کلمات			کلمه+POS+ریشه			کلمه+POS+ریشه		
اندازهٔ پنجره	۵ توکن			۳ توکن			۵ توکن		
برچسب/ معیار	F	R	P	F	R	P	F	R	P
B-cause	۶۵	۵۷	۷۵/۵	۶۸	۶۹	۷۵	۶۸	۶۵	۷۲
I-cause	۵۹	۵۹	۵۹/۸	۶۳	۶۵	۶۲	۶۶	۷۱	۶۱
B-effect	۷۴	۶۸	۸۰	۷۱/۴	۶۹	۸۲	۷۷	۷۲	۸۲
I-effect	۷۶	۷۳	۷۹	۷۵	۷۱	۷۹	۷۹	۷۴	۸۴
B-mark	۸۸	۸۴	۹۳	۸۹	۸۶	۹۲	۹۴	۹۲	۹۶
I-mark	۷۵	۶۹	۸۲	۷۱,۴	۷۷	۶۶	۷۴	۷۷	۷۱
O	۷۳	۷۶	۶۹	۷۵	۷۸	۷۲	۷۳	۷۵	۷۱
میانگین	۷۳	۶۹	۷۷	۷۴	۷۳	۷۵	۷۶	۷۵	۷۷

تغییر اندازهٔ پنجرهٔ مجاورت در روش طبقه‌بندی CRF به‌طور مستقیم بر عملکرد سیستم تأثیر می‌گذارد. بهترین نتیجهٔ ما از طریق پنجرهٔ ۵ توکن به‌دست می‌آید. ما آزمایش‌های خود را با POS و بدون آن و ویژگی‌های پایه تکرار کردیم. همان‌طور که در جدول ۱۰، مشاهده می‌شود، افزودن ویژگی POS و ریشه (افزودن ویژگی ریشه به‌تنهایی به میزان ۰/۵ واحد بهبود ایجاد می‌کند)، نتایج کل را بر اساس معیار F بهبود داده و عملکرد سیستم را به‌طور میانگین تا ۳ درصد در معیار F بهبود می‌بخشد و بهترین عملکرد برای پنجرهٔ ۵ توکنی به‌دست می‌آید. برای نمایش بهتر دو نمودار در شکل‌های ۵ و ۶ برای مقایسهٔ نتایج سیستم CRF با توجه به اندازهٔ پنجره و سایر ویژگی‌ها ارائه شده است.



شکل ۵. ارزیابی عملکرد سیستم CRF (معیار F برای هر برچسب با پنجرهٔ سایز ۵ توکن و ویژگی‌های متفاوت)



شکل ٦. ارزیابی عملکرد سیستم CRF (معیار F برای هر برچسب با پنجره‌های متفاوت)

اگر بخواهیم بخش یافته‌ها را به‌طور خلاصه مرور کنیم، در این بخش آزمایشات مختلفی برای بررسی کارایی روش‌های مختلف یادگیری ماشین و یادگیری عمیق در راستای ایجاد یک سیستم شناسایی مرز عناصر علّی با استفاده از پیکرهٔ PerCause انجام

شد. سرانجام، CRF با اندازهٔ دادگان فعلی بهترین عملکرد کلی را در بین تمام سیستم‌های آزمایش‌شده دارد و این نتیجه نیز با در نظر گرفتن خود کلمه، ریشهٔ آن و برچسب POS و با استفاده از پنجرهٔ ۵توکنی به‌دست می‌آید.

۵. بحث و نتیجه‌گیری

همان‌طور که در بخش قبل نشان داده شد، روش CRF با استفاده از PerCause بهترین نتایج کل بر اساس معیار F را در بین سیستم‌های پیاده‌شده دارد. الگوریتم CRF بر اساس احتمال شرطی و نظریهٔ گراف است و بر خلاف بسیاری از روش‌های دیگر، تکنیک تصادفی شرطی برای هر نمونه، به‌جای در نظر گرفتن نمونه به‌تنهایی، برچسب‌های کلمات قبلی یا بعدی را نیز در نظر می‌گیرد. به نظر می‌رسد که این موضوع نقطهٔ قوت این روش در بین روش‌های دیگر است. همچنین، نتایج تجربی نشان می‌دهد که در سیستم CRF، اندازهٔ پنجره و ویژگی‌های پایهٔ POS و ریشهٔ کلمه پارامترهای مؤثری هستند که عملکرد سیستم را بهبود می‌بخشند.

افزون بر این، انتظار می‌رفت که روش یادگیری عمیق بهترین عملکرد کلی را داشته باشد که احتمالاً به‌دلیل مجموعه دادهٔ ناکافی، الگوریتم به‌رغم داشتن بهترین صحت (۹۱/۴ درصد) برتر از سایرین نبود. آزمایش‌ها نشان می‌دهند که اندازهٔ داده‌های آموزشی تأثیر آشکاری بر نتایج در سیستم یادگیری عمیق دارد. بدین صورت که پس از افزایش اندازهٔ داده‌ها (تقریباً دو برابر اندازهٔ اولیه را به مجموعه اول اضافه کردیم)، مقدار معیار F سیستم تا ۱۳ درصد افزایش یافته است. اما این موضوع بر عملکرد CRF تأثیر چندانی نمی‌گذارد. پس، به‌صورت خلاصه، برای اندازهٔ داده‌های کوچک بهتر است از طبقه‌بندی‌کنندهٔ CRF استفاده شود، اما یادگیری عمیق در اندازهٔ داده‌های در مقیاس بزرگ انتخاب بهتری است.

نکتهٔ بعد این است که مجموعهٔ آموزشی که به‌طور خودکار ایجاد شده، ظاهراً شبکهٔ عمیق را گمراه کرده و به نتایج بهتری منجر نشد. شاید حرکت به سمت نظارت ضعیف این موضوع را مدیریت کند.

از آنجا که اندازهٔ مجموعهٔ دادگان علّی تولیدشده فعلی کوچک است، این مجموعه به‌صورت درشت‌دانه تهیه شده که در آن همهٔ انواع علّیت در یک برچسب کلی از «رابطهٔ علّی» ادغام شده‌اند. در آینده، مجموعهٔ دادهٔ بزرگ‌تر با برچسب‌های ریزدانه مانند «هدف» و «انگیزه» خواهیم داشت.

در این مقاله، فرایند تولید یک پیکرهٔ برچسب‌گذاری‌شدهٔ علّیت برای فارسی ارائه شد. افزون بر این، از چند روش یادگیری ماشین و یادگیری عمیق برای آموزش مدل برای تعیین مرزهای عناصر علّی با استفاده از این مجموعه داده استفاده شد. در میان الگوریتم‌های یادگیری ماشین آزمایش‌شده، روش CRF با معیار Fبرابر ۷۶، احتمالاً به‌دلیل ویژگی مدل‌سازی بافتار و نه نمونه تنها، بهترین عملکرد کل را دارد. در ضمن، دو معماری یادگیری عمیق bi- LSTM+CRFو مبتنی بر «برت» نیز ارزیابی شد که bi-LSTM+CRF بهترین دقت را در میان سیستم‌های عمیق آزمایش‌شده نشان داد؛ هرچند، به‌دلیل اندازهٔ ناکافی داده، نتیجهٔ کل با معیار F مطابق انتظار نبود و داده‌های تولیدشده به‌طور خودکار عملکرد خوبی نیز نداشتند.

سرانجام این‌که با استفاده از سیستم فعلی می‌توان با دقت مناسبی مرز عناصر علّی را مشخص و از جملات کاندید زوج‌های علّی را استخراج نمود. این امر که هدف کاربردی این پژوهش است، می‌تواند در بسیاری از وظایف پردازش زبان در فارسی از قبیل ساخت منبع دانش و استخراج رویدادهای متن کمک‌کننده باشد.

۶. پیشنهادات پژوهش و کارهای آینده

به‌عنوان کار آینده، افزودن پس‌پردازش و اصلاح دستی مجموعه داده‌های تولیدشدهٔ خودکار یا حرکت به سمت روش‌های بانظارت ضعیف احتمالاً نتایج را بهبود می‌بخشد. افزون بر این، افزودن ویژگی‌های دیگری مانند قطعه‌بندی یا نقش معنایی ممکن است سیستم CRF را بهبود بخشد.

References

Bijankhan, Mahmoud. 2004. The role of the corpus in writing a grammar: An introduction to a software. *Iranian Linguistic Journal 19 (2):* 48-67.

Blanco, Eduardo, Nuria Castell, and Dan Moldovan. 2008. Causal relation extraction. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Morocco.

Chang, Du-Seong, and Key-Sun Choi. 2004. Causal relation extraction using cue phrase and lexical pair probabilities. In *International Conference on Natural Language Processing*. pp. 61-70. Springer, Berlin, Heidelberg.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*37-46 :(1) *20* .

Dasgupta, Tirthankar, Rupsa Saha, Lipika Dey, and Abir Naskar. 2018. Automatic extraction of causal relations from text using linguistically informed deep neural networks. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 306-316. Melbourne, Australia.

Deleger, Louise, Qi Li, Todd Lingren, Megan Kaiser, and Katalin Molnar. 2012. Building gold standard corpora for medical natural language processing tasks. In *AMIA Annual Symposium Proceedings*, vol. 2012, p. 144. Chicago, USA: American Medical Informatics Association.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv: 1810.04805*.

Dunietz, Jesse, Lori Levin, and Jaime G. Carbonell. 2015. "Annotating causal language using corpus lexicography of constructions." In *Proceedings of the 9th Linguistic Annotation Workshop*, pp. 188-196.

Dunietz, Jesse, Lori Levin, and Jaime G. Carbonell. 2017." The BECauSE corpus 2.0: Annotating causality and overlapping relations." In *Proceedings of the 11th Linguistic Annotation Workshop*, pp. 95-104.

Farahani, Mehrdad, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021. "Parsbert: Transformer-based model for persian language understanding." *Neural Processing Letters* 53, no. 6: 3831-3847.

Garcia, Daniela. 1997. "COATIS, an NLP system to locate expressions of actions connected by causality links." In *International Conference on Knowledge Engineering and Knowledge Management*, pp. 347-352. Springer, Berlin, Heidelberg.

Girju, Roxana. 2003. "Automatic detection of causal relations for question answering." In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering*, pp. 76-83.

Goyal, Archana, Kumar Manish, and Gupta Vishal. 2017 ."Named entity recognition: applications, approaches and challenges." *International Journal of Advance Research in Science and Engineering 35 (5):* 482-489.

Goyal, Archana, Vishal Gupta, and Manish Kumar. 2018 ."Recent named entity recognition and classification techniques: a systematic review." *Computer Science Review 29:21-43.*

Green, Annette M. 1997. "Kappa statistics for multiple raters using categorical classifications." In *Proceedings of the 22nd annual SAS User Group International conference*, vol. 2, p. 4.

Gurulingappa, Harsha, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012 ."Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports." *Journal of biomedical informatics* 45, no. 5: 885-892.

Hashimoto, Chikara. .2019 "Weakly supervised multilingual causality extraction from Wikipedia." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2988-2999.

Huang, Zhiheng, Wei Xu, and Kai Yu. 2015. "Bidirectional LSTM-CRF models for sequence tagging." *arXiv preprint arXiv: 1508.01991*.

Karimi, Sarvnaz, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics 55: 73-81.*

Khoo, Christopher SG, Chan Syin, and Niu Yun. 2000. Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of the 38th annual meeting of the association for computational linguistics*, pp. pp. 336-343. Hong Kong.

Leaman, Robert, Christopher Miller, and Graciela Gonzalez. 2009. Enabling recognition of diseases in biomedical text with machine learning: corpus and benchmark. In *Proceedings of the 2009 Symposium on Languages in Biology and Medicine*, vol. 82, no. 9. Jeju Island, South Korea.

Luo, Zhiyi, Yuchen Sha, Kenny Q. Zhu, Seung-won Hwang, and Zhongyuan Wang. 2016. Commonsense causal reasoning between short texts. In *Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Cape Town, South Africa.

McCallum, Andrew, and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of CoNLL*. *Edmonton, Canada.*

Mihăilă, Claudiu, Tomoko Ohta, Sampo Pyysalo, and Sophia Ananiadou. 2013 .BioCause: Annotating and analysing causality in the biomedical domain. *BMC bioinformatics 14 (1): 1-18.*

Mirza, Paramita. 2014. Extracting temporal and causal relations between events. In *Proceedings of the ACL 2014 Student Research Workshop*, pp. 10-17. *Baltimore, Maryland, USA.*

Mirza, Paramita, and Sara Tonelli. 2016. Catena: Causal and temporal relation extraction from natural language texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, *Osaka, Japan: Technical Papers*, pp. 64-75.

Mostafazadeh, Nasrin, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pp. 51-61. *San Diego, California.*

Ning, Qiang, Zhili Feng, Hao Wu, and Dan Roth. 2019 .Joint reasoning for temporal and causal relations. *arXiv preprint arXiv: 1906.04941*.

Ramshaw, Lance A., and Mitchell P. Marcus. 1995. *Text chunking using transformation-based learning*. *Natural language processing using very large corpora.* Dordrecht: Springer.

Rehbein, Ines, and Josef Ruppenhofer. 2020. A new resource for German causal language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 5968-5977. Pharo.

Sadek, Jawad, and Farid Meziane. 2018. Building a causation annotated corpus: the Salford Arabic Causal Bank-proclitics. In *11th Edition of the Language Resources and Evaluation Conference*. Miyazaki Japan.

Sadek, Jawad, and Farid Meziane. 2018. Learning causality for Arabic-proclitics. *Procedia computer science 142: 141-149.*

Schneider, Nathan, Jena D. Hwang, Vivek Srikumar, Meredith Green, Kathryn Conger, Tim O'Gorman, and Martha Palmer. 2016. A corpus of preposition supersenses in English web reviews. *arXiv preprint arXiv: 1605.02257*.

Schuler, Karin Kipper. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Pennsylvania: University of Pennsylvania.

Shamsfard, Mehrnoush, Hoda Sadat Jafari, and Mahdi Ilbeygi. 2010. STeP-1: A Set of Fundamental Tools for Persian Text Processing. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10) 2010. May. Malta.

زینب رحیمی

متولد سال ۱۳۶۶، دارای مدرک کارشناسی ارشد فناوری اطلاعات از دانشگاه صنعتی امیرکبیر است. ایشان هم‌اکنون دانشجوی دکتری در رشتهٔ مهندسی کامپیوتر، گرایش هوش مصنوعی در دانشگاه شهید بهشتی است.

پردازش زبان طبیعی، داده‌کاوی و یادگیری ماشین از جمله علایق پژوهشی وی است.

مهرنوش شمس‌فرد

دارای مدرک دکتری از دانشگاه صنعتی امیرکبیر است. ایشان هم‌اکنون دانشیار دانشکدهٔ مهندسی و علوم کامپیوتر و سرپرست آزمایشگاه پردازش زبان طبیعی آن دانشگاه است.

پردازش زبان طبیعی با تأکید بر زبان فارسی، معناشناسی رایانشی، مهندسی دانش، و وب معنایی از جمله علایق پژوهشی وی است.