# کشف روابط علّی در متون فارسی با تکنیک‌های یادگیری ژرف

**حمیدرضا اکبری**

*hakbari@gmail.com*

**امین سعیدی**

*Amin.saeidi.1997@gmail.com*

**زهرا فضلی**

*z.fazli124@yahoo.com*

**مهدی فرح‌بخش**

*Mahdi124710@yahoo.com*

**علیرضا مسلمی‌حقیقی**

*Ali\_m110022@yahoo.com*

**چکیده**

*با افزایش حجم داده ها تشخیص و استخراج هوشمند روابط بین کلمات و عبارت ها در متون کاربردهای روز افزون زیادی داشته است. یکی از روابط مهم بین کلمات و عبارتها کشف روابط* علّی *در متون و جملات است. کارآمدی شبکه های عصبی در حل مسائل یادگیری هوشمند، انتخاب این راهکار را به عنوان یکی از متداول ترین روش های یافتن روایط* علّی *در جملات قرار داده است. در این تحقیق روش های متعددی مبتنی بر شبکه های عصبی برای استخراج روابط* علّی *بررسی شده است. بکارگیری از شبکه های بازگشتی، مدل های مبتی بر ترانسفورمر و مدل های مبتنی بر بازنمایی کلمات از جمله روشهایی است که در این تحقیق بررسی شده است. از آنجاییکه موضوع تحقیق یک مسئله بانظارت است از منابع مختلف اطلاعاتی داده های اولیه جمع آوری شده و به صورت دستی برچسب های مورد نظر اعمال شده است. هرچند مدلهای مبتنی بر ترانسفورمر از قابلیتهای بازنمایی مبتنی بر سیاق و مکانیزم توجه بهره برده و کارایی قابل قبولی دارند اما مدلهای مبتنی بر شبکه های بازگشتی با بهره گیری از بازنمایی کلمات نیز می توانند از جایگاه ویژه ی به جهت سادگی و سبک بودن مدل برخوردار باشند. افزودن مکانیزم توجه به این مدل ها می تواند زمینه تحقیق بعدی و تکمیل این تحقیق در نظر گرفته شود.*

**کلمات کلیدی:** کشف روابط علت و معلولی[[1]](#footnote-1)- شبکه‌های عصبی بازگشتی[[2]](#footnote-2)- مدل های مبتنی بر ترانسفورمر[[3]](#footnote-3) - بازنمایی مبتنی بر سیاق[[4]](#footnote-4) – مکانیزم توجه[[5]](#footnote-5)- یادگیری ژرف[[6]](#footnote-6)

**۱- مقدمه**

استخراج روابط علت‌و‌معلولی از متن، یک مساله چالش‌برانگیز در پردازش زبان‌های طبیعی[[7]](#footnote-7) است که در سال‌های اخیر مورد توجه زیادی قرار گرفته است. این مساله که خود زیرمجموعه مساله بزرگتری برای استخراج انواع روابط واژگانی[[8]](#footnote-8) در متن است، بررسی راهکارهای موجود و چالش های هر راهکار را بیش از بیش مورد توجه قرار می دهد. بررسی های اولیه نشان از تحقیقات گسترده استخراج روابط علی در متون به زبان انگلیسی می باشد و تحقیقات پیرامون این مسئله در زبان فارسی کمتر مشاهده می‌شود. ازهمین‌رو ابتدا به رویکردهای موجود در زبان انگلیسی پرداخته و سپس به زبان فارسی خواهیم پرداخت، هرچند تفاوت‌های مشاهده شده چندان نبوده و بیشتر به داده‌های آموزش، جمع‌اوری، برچسب‌گذاری و یکسری تفاوت‌های زبانی در قرارگیری کلمات است.

به طور کلی، بررسی روابط علت و معلولی به دو روش کلی انجام می‌شود، روش‌های مبتنی بر قانون و روش‌های عصبی. یکی از روش‌های مبتنی بر قانون روش ارایه شده توسط خو و همکاران (۲۰۰۰) است، آنها روشی را برای شناسایی و استخراج اطلاعات علت و معلولی ایجاد می کنند که در چکیده های پزشکی در پایگاه داده Medline بیان شده است. برای این منظور مجموعه ای از الگوهای گرافیکی ساخته می‌شود که نشان دهنده وجود رابطه علّی در جملات است، و اینکه کدام قسمت جمله بیانگر علت و کدام قسمت نشان دهنده معلول است. الگوها با درخت های تجزیه نحوی جملات مطابقت داده می شوند و قسمت هایی از درخت تجزیه که با شکاف های موجود در الگوها مطابقت دارند به عنوان علت یا معلول استخراج می شوند. یک کار دیگر در این حوزه کار لیو و همکاران(۲۰۱۶) است، آنها به‌طور خودکار شبکه‌ای از اصطلاحات علی را از یک مجموعه وب بزرگ جمع‌آوری می‌کنند و به کمک آن یک معیار جدید و موثر برای مدل‌سازی مناسب علیت بین عبارت‌ها پیشنهاد می‌کنند.

از جمله کارهایی که از روش‌های عصبی استفاده می‌کند می توان به مدل‌های یادگیری عمیق مبتنی بر شبکه‌های کانولوشن[[9]](#footnote-9)، شبکه‌های بازگشتی، و مکانیسم‌های توجه اشاره کرد. در بخش استخراج نوع رابطه، ژنگ و همکاران (2015) یک رویکرد مبتنی بر شبکه عصبی برای استخراج رابطه علت و معلولی مبتنی بر شبکه های کانولوشن پیشنهاد کردند. این مدل از یک لایه کانولوشن برای استخراج ویژگی های متنی و یک لایه Softmax برای طبقه بندی توکن ها استفاده می­کند. نویسندگان رویکرد خود را بر روی یک مجموعه داده معیار ارزیابی کردند و نتایج رقابتی را گزارش کردند. لیو و همکاران (2019) یک مدل شبکه عصبی انتها به انتها[[10]](#footnote-10) ارائه کردند که از مکانیزم توجه دوگانه برای به‌تصویر‌کشیدن تعاملات بین موجودیت ها[[11]](#footnote-11) در جمله برای استخراج رابطه استفاده می‌کند. رویکرد آنها به نتایج پیشرفته‌ای در چندین مجموعه داده معیار در تشخیص رابطه علّی منجر شده است. نگوین و همکاران (2020) یک مدل استخراج مشترک برای رویدادها و روابط علّی آن‌ها از متن با استفاده از رویکرد یادگیری چند‌وظیفه­ای[[12]](#footnote-12) پیشنهاد کرد. مدل آن‌ها ترکیبی از لایه های کانولوشن و بازگشتی برای بازنمایی متن ورودی طراحی شده است و نویسندگان نشان دادند که یادگیری چند وظیفه‌ای می‌تواند عملکرد مدل را هم در استخراج رویداد و هم در استخراج رابطه علی بهبود بخشد. لی و همکاران (2021) یک رویکرد مبتنی بر شبکه عصبی را برای استخراج رابطه علی ارائه کرد که هم از اطلاعات نحوی و هم از اطلاعات معنایی استفاده می‌کند. این مدل از یک شبکه عصبی بازگشتی برای گرفتن ساختار نحوی متن ورودی و لایه کانولوشن برای استخراج ویژگی‌های معنایی استفاده می‌کند. نویسندگان رویکرد خود را بر روی یک مجموعه داده معیار ارزیابی کردند و نتایج رقابتی را گزارش کردند.

مطالعات دیگر نیز استفاده از شبکه‌های عصبی را برای استخراج رابطه علت و معلولی، از جمله رویکردهای ترکیبی که روش‌های مبتنی بر قانون را با شبکه‌های عصبی ترکیب می‌کنند، مورد بررسی قرار داده‌اند. به عنوان مثال، وانگ و همکاران. (2019) یک رویکرد ترکیبی پیشنهاد کرد که از مجموعه ای از قوانین دست‌ساز برای استخراج روابط علت و معلولی بالقوه از متن ورودی استفاده می‌کند، که سپس برای آموزش از یک مدل مبتنی بر کانولوشن برای پیش‌بینی روابط علت و معلولی نهایی استفاده می‌شود. علاوه‌ بر‌ این، تلاش‌هایی برای بهبود عملکرد مدل‌های مبتنی بر شبکه عصبی برای استخراج رابطه علت و معلولی از طریق تکنیک‌های مختلف صورت گرفته‌است. به‌عنوان مثال، ژانگ و همکاران. (2019) روشی را برای ترکیب دانش خارجی در یک مدل مبتنی بر شبکه عصبی با استفاده از تعبیه‌های کلمه از‌پیش‌آموزش‌داده‌شده بر روی مجموعه بزرگی از متن پیشنهاد کرد. رویکرد آن‌ها به نتایج پیشرفته‌ای در چندین مجموعه داده معیار دست‌یافت. رویکرد دیگر برای بهبود عملکرد مدل‌های مبتنی بر شبکه عصبی برای استخراج رابطه علت‌و‌معلولی از طریق یادگیری انتقالی است. گوو و همکاران (2019) یک رویکرد یادگیری انتقالی[[13]](#footnote-13) را پیشنهاد کرد که از مدل‌های زبانی از‌پیش‌آموزش‌دیده[[14]](#footnote-14) برای بهبود عملکرد یک مدل مبتنی بر شبکه عصبی برای استخراج رابطه، از جمله روابط علت و معلولی استفاده می‌کند.

از معدود مطالعات صورت‌گرفته در زبان فارسی رحیمی، شمس‌فرد (۲۰۲۱) است که با تجمیع داده‌های دستی غیرقابل‌دسترس برای عموم، پژوهش خود را انجام داده‌اند. آن‌ها با انجام برچسب‌گذاری بایو[[15]](#footnote-15) و سپس اعمال مدلهای متفاوت خروجی‌ها را مقایسه و نتایج رقابتی را گزارش نموده‌اند. مطالعه آن‌ها قابل مقایسه با سایر کارهای صورت‌گرفته مانند این پژوهش نخواهد بود چراکه جزییات مطالعه برای بررسی دقیق تفاوت و شباهت‌های بین مدل‌ها وجود ندارد.

**۲- جمع‌آوری دادگان**

از آنجاییکه مسئله یافتن روابط علت معلولی یک مسئله بانظارت[[16]](#footnote-16) است جمع آوری و تهیه دادگان مناسب از اهمین ویژه ی برخوردار است. بررسی انواع داده های موجود جهت استفاده از این تحقیق اولین گام در بخش جمع آوری دادگان می­باشد. به عنوان مثال، مجموعه دادگان The SemEval 2010- Task 8 (هندریکس و همکاران ۲۰۱۰) که برای زبان فارسی نیز توسعه یافته به جهت اینکه تنها رابطه علت‌ومعلولی را شامل نمی‌شود، مجموعه محدودی از داده ها را ارائه می نماید. پیکره Causal TimeBank (میرزا و تونلی ۲۰۱۶) نیز دادگان دیگری برای زبان فارسی بوده اما در دسترس عموم قرار نگرفته است. از‌همین‌رو، توسعه مجزای دادگان یکی از الزامات تحقیق خواهد بود. در ادامه به جزئیات منابع داده­ی و شیوه تولید داده ها اشاره شده است.

شایان ذکر است که مجموع داده هایی که توسط تیم تحقیق به صورت دستی برچسب گذاری شده **4135** جمله علت و معلولی و دادگان آزمایشگاه پردازش زبان طبیعی دانشگاه شهید بهشتی **4446** جمله و در مجموع **8581** جمله می­باشد.

**۲-۱ ویکی پدیا فارسی**

یکی از منابع داده­ی استفاده از محتویات ویکی­پدیا فارسی[[17]](#footnote-17) می باشد. اندازه اطلاعات مذکور بیش از 780مگابایت بوده و اطلاعات فارسی موجود تا تاریخ مارچ 2020 گردآوری شده است. برای استخراج روابط علی از کلمات کلیدی متناسب با این حوزه استفاده شده است. برخی از عبارت های بکارگرفته شده عبارتند از : معلول،علت، به دلیل، چون، به جهت، ارتباط، بلکه و غیره. برای استخراج جملات احتمالی علت و معلول، ابتدا از متن های موجود جملات را استخراج کرده و سپس جملاتی که عبارت های اشاره شده در آن وجود دارد جهت برچسب زنی دستی استخراج می شود.

حدود 10.000 جمله از متن استخراج شده و بین اعضای تیم تحقیق جهت برچسب زنی تقسیم شده است. اعضا تیم تحقیق سه نوع برچسب را بر روی جملات تعیین کرده اند. برچسب ها عبارتند از مارکر، علت و معلول. مارکر به عبارتهایی اشاره دارد که رابط روابط علت و معلولی است. نمونه ی از مارکرها در بالا اشاره شده است. مجموع داده های برچسب خورده از این دادگان نزدیک به 2000 جمله می­باشد.

برای برچسب زنی بر روی جملات حتی الامکان ملاحظات ذیل در برچسب زنی رعایت شده است.

* برجسب ها به محدوده دقیق عبارت های مارکر، علت و معلول اشاره داشته باشد
* به جهت تعمیم پذیری محدوده های برچسب ها به گونه تعیین شود که دربرگیرنده چارچوب عمومی روابط باشد
* امکان عدم وجود مارکر در روابط علت معلولی در نظر گرفته شده است.
* امکان وجود بیش از یک علت و یا معلول نیز در جملات در نظر گرفته شده است.

**۲-۲ داده های** PERLEX

برای استخراج روابط از مجموعه دادگان فارسی یک مجموعه داده با نام PERLEX[[18]](#footnote-18) بررسی شده است. این مجموعه داده اختصاصی حوزه روابط علت و معلولی نبوده و به صورت عمومی تمامی روابط را شامل می شود. در بخش آموزش حدود 32000جمله برچسب خورده است. 853 جمله از این موارد به روابط علت و معلولی اشاره دارد. توجه به این مهم ضروری است که داده های این مجموعه برچسب مارکر ندارند. در عمده مدل ها وجود مارکر در هر جمله اختیاری فرض شده است.

**۲-3 داده های** Causal Bank

یکی دیگر از منابعی که برای این منظور مورد بررسی قرار گرفت، استفاده از منابع انگلیسی و ترجمه‌ی آن‌ها به کمک کتابخانه‌ی مترجم زبان گوگل در پایتون است. برای این منظور از مجموعه‌ي دادگان causal bank[[19]](#footnote-19) استفاده شد. این منبع شامل ۲۷۰ میلیون داده‌ی علت و معلولی به زبان انگلیسی است. کار روی بخشی از این داده انجام شد و در نهایت مشاهده گردید، جملات تولیدی در مواقعی که در جمله‌ی عدد به کار رفته است، به خوبی ترجمه نمی‌شود، برای همین ، کار با حذف این جملات تکرار شد. در نهایت مشاهده گردید که باز هم به دلیل ترجمه‌ي نامناسب و عدم سازگاری جملات انگلیسی با فرهنگ ما و مناسب نبودن معنایی جملات تولیدی در زبان فارسی، داده‌ی تولیدی قابلیت تزریق به مدل را ندارد و بنابراین کنار گذاشته شد.

**۲-4 دادگان پیکره علییت آزمایشگاه پردازش زبان های طبیعی دانشگاه شهیدبهشتی[[20]](#footnote-20)**

با توجه به انجام تحقیقات مشابه در آزمایشگاه پردازش زبان های طبیعی دانشگاه شهید بهشتی، مکاتباتی با سرکارخانم دکتر مهرنوش شمس فرد، سرپرست آزمایشگاه صورت گرفت و ایشان مجموعه دادگان را در اختیار این تحقیق قرار دادند. پیکره برچسب گذاری شده علیت توسعه یافته در آزمایشگاه پردازش زبان طبیعی دانشگاه شهیدبهشتی، تقریباً از 130.000 توکن و 5128 رابطه علی در قالب 4446 جمله تشکیل شده است. پیکره خام اولیه که از میان متون دو پیکره 1) بیجن‌خان و 2) کتاب (یک مجموعه عمومی متشکل از 1.8 میلیون کلمه از 10 رمان و کتاب عمومی فارسی) انتخاب شده، به صورت دستی برچسب گذاری شده است.

همانطور که مفهوم علیت اقتضا می‌کند، مجموعه برچسب‌ها حاوی 3 برچسب "علت" ، "معلول" و "نشانه علیت" است و پیکره برچسب خورده در قالب برچسب های IOB آماده شده است. از ویژگی‌های این پیکره می‌توان به عدم الزام به توالی کلمات در یک رکن رابطه علّی اشاره کرد. چون گاهی یک عنصر علّی (مثلا علت) ممکن است شامل دو یا چند قسمت باشد که لزوما متوالی نیستند. همچنین الگوهای مورد نظر برای استخراج جملات کاندید برای برچسب‌گذاری محدودیت گرامری ندارند. برای مشاهده جزئیات پیکره و ارجاع به آن، به مقاله[[21]](#footnote-21) زیر رجوع بفرمایید.

**۲-5 ساخت‌داده با ChatGPT**

یکی دیگر از روش‌های مورد استفاده برای ساخت داده، استفاده از ChatGPT به‌عنوان جمله‌ساز بوده است. از هر ۱۰۰۰ داده آماده‌شده ۶۰۰ تا توسط این نرم‌افزار ساخته شده و بعد از ترجمه به فارسی و بازنویسی مجدد برای گنجاندن ۱۰۰ الگوی ازپیش‌مشخص‌شده (نشانه‌گرها) مورد برچسب‌گذاری قرار گرفته‌است. برای داشتن تنوع زیاد نیز هر ۲۰ تا ۳۰ جمله با یک عنوان مشخص از نرم‌افزار تولید شده‌اند. برای مثال، عنوان‌های اقتصادی، سیاسی، ورزشی، سلامت و ... مورد استفاده قرار گرفته‌اند. داده‌های تست نیز به همین عنوان تولید شده و ۲۰۰ تا از هر ۱۰۰۰ جمله نیز برای این روش بدون رابطه علیت ایجاد شده تا به استحکام بیشتر شبکه کمک کند. ۲۰۰ داده آخر نیز از این ۱۰۰۰‌تایی‌ها به‌صورت دستی تولید شده‌اند.

**۳- مدل های یادگیری**

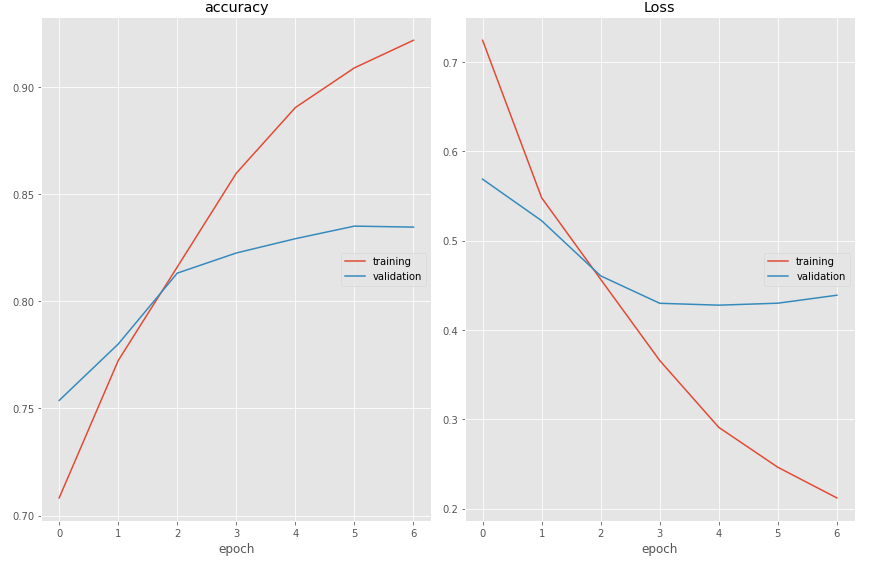
در این بخش از تحقیق به بررسی انواع مدل های یادگیری اشاره می شود. در هر مدل ابتدا به معماری مدل و پارامترهای آن اشاره شده و سپس نتایج حاصل از آموزش و ارزیابی مدل تشریح می شود. در انتهای هر مدل نیز نقاط قوت و ضعف هر مدل اشاره می شود.

**3-۱ شناسایی روابط علت و معلولی با شبکه های بازگشتی**

برای پیاده سازی این مدل ابتدا لازم است تا داده های برچسب خورده نرمال شده و توکن هایشان جدا شود. سپس مراحل آماده سازی داده انجام می‌شود تا دادگان برای خوراک مدل مهیا شوند. سپس دادگان به سه بخش داده های آموزش، تست و اعتبار سنجی تقسیم می­شوند و به مدلی با جزئیات زیر داده می‌شوند تا یادگیری اتفاق بیافتد. لازم به ذکر است که در ابتدا دادگان توسط تابع tensorflow.keras.preprocessing.sequence.pad\_sequences به برداری با بعد تعداد جملات در 40 تبدیل می­شوند و سپس به شبکه ای داده می­شوند که در لایه اول جاسازی توسط tensorflow.keras.layers.Embedding انجام می­شود که هرجمله را به برداری با بعد 50 تبدیل می­کند. سپس در لایه بعدی یک لایه dropout یک بعدی اضافه می­شود تا این لایه به طور تصادفی واحدهای ورودی را با فرکانس نرخ در هر مرحله در طول زمان آموزش روی 0 تنظیم کند و در نهایت از overfit شدن مدل جلوگیری کند. لایه بعدی شبکه یک lstm دو طرفه است. تابع فعالسازی لایه آخر مدل نیز softmax است تا احتمالات را به برچسب ها تخصیص دهد. مدل با سایز batch 32 و در 6 epoch آموزش می­بیند که تغییرات دقت و تابع ضرر در تصویر زیر مشاهده می­شود. به طور خلاصه معماری مدل از لایه استاندارد Embedding، لایه GRU، لایه TimeDistributed و لایه SpatialDropout برای منظم سازی استفاده می کند.

همجنین نتایج برای دادگان تست نیز دقت 83 درصدی را برای پیش بینی درست هر برچسب نشان می­دهد. در آخر روی چند داده‌ی آزمایشی، پیش‌بینی کردن تگ‌ها انجام می­شود. تگ‌ها در این قسمت به این صورت استفاده شده‌است:

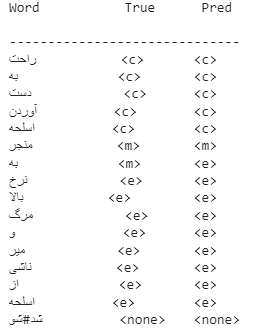
* n : سایر کلمات
* c: علت



شکل ۱- نمودار یادگیری و اعتبارسنجی مدل اول

* e : معلول
* m: مارکر

نتایج پیش‌بینی بر روی یک جمله در ادامه قرار داده شده است.



شکل ۲- نمونه پیش‌بینی مدل اول

**3-2 شناسایی روابط علت و معلولی با شبکه های بازگشتی و بازنمایی FastText**

برای این پیاده سازی پس از نرمال سازی و جدا کردن توکن‌ها به کمک fasttext جاسازی می‌شوند. کل جملات به صورت داده‌های آموزش، اعتبار سنج و تست تقسیم می‌شوند. یک شبکه‌ی lstm یک لایه و یک طرفه ایجاد کرده و سایز batch را ۶۴ در نظر گرفته و در سی epoch آموزش انجام داده می‌شود. یک بار دیگر کار را با شبکه ی دو لایه و دو طرفه تکرار می‌کنیم، چون نتایج مرحله‌ی اول بهتر بود فقط نتایج مرحله‌ي اول گزارش می‌شود. جدول زیر نتایج را روی اجرای ایپاک‌ها را نشان می‌دهد :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| accuracy | f1 | recall | Validation loss | Training loss |  |
| 64.595933 | 0.612506 | 0.588132 | 0.846877  جدول ۱- خروجی معیارهای آموزشی مدل دوم | 0.822449 | Epoch30 |

نتیجه ارزیابی روی داده های تست نیز به شرح ذیل می­باشد:

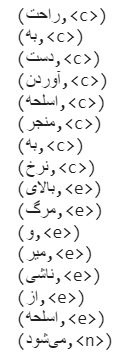
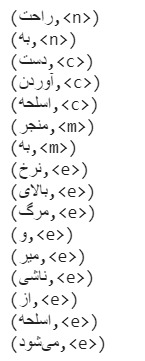


در نهایت روی چند داده‌ی آزمایشی، پیش‌بینی کردن تگ‌ها را انجام می‌دهیم. تگ‌ها در این قسمت به این صورت استفاده شده‌است:

* n : سایر کلمات
* c: علت
* e : معلول
* m: مارکر

نتایج پیش‌بینی روی دو جمله در ادامه قرار داده شده است، تصویر سمت راست مربوط به شبکه‌ی تک لایه و تصویر سمت چپ مربوط به شبکه‌ی دوطرفه و دو لایه است. نتایج کاملا دقیق نیست ولی با توجه به تعداد داده‌ی تست، پیش‌بینی قابل قبولی می‌باشد.

شکل ۳- نمونه پیش‌بینی مدل دوم



**3-3 شناسایی روابط علت و معلولی با مدل زبانی BERT**

مدل برت یک مدل زبانی کارآمد مبتنی بر معماری ترانسفورمر می باشد. در این مدل بازنمایی سیاق محور کلمات بکارگرفته شده است. یکی از مزیتهای این مدل استفاده از مکانیزم توجه چند سر[[22]](#footnote-22) برای بررسی میزان توجه عبارتهای یک متن می­باشد. برای حل مسئله استخراج روابط علت و معلول از یک قابلیت برت در مسئله طبقه بندی توکن ها[[23]](#footnote-23) استفاده شده است. در این مدل از 12 لایه که هر لایه مشتمل بر موارد ذیل است:

* BertAttention
* BertSelfOutput
* BertIntermediate
* BertOutput

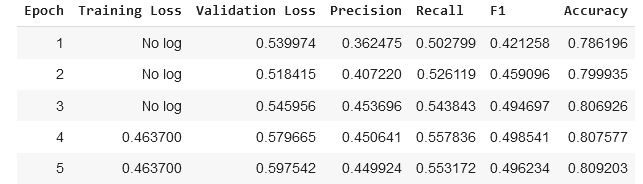
در لایه انتهای از یک شبکه طبقه بندی جهت تشخیص برچسب کلمات استفاده شده است. ابعاد ورودی این لایه بردار بازنمایی کلمات در مدل برت بوده و ابعاد خروجی تعداد برچسب های کلمات می باشد. برچسب ها به شرح ذیل می­باشند.

* O : سایر کلمات
* B-CAUSE : شروع علت
* I-CAUSE : کلمه میانی علت
* B-EFFECT : شروع علت
* I- EFFECT: کلمه میانی علت
* B- MARKER: شروع نشانگر
* I- MARKER: کلمه میانی نشانگر

مجموعه داده های آموزش، اعتبارسنجی و تست به ترتیب به نسبت 80-10-10 تقسیم شده است.

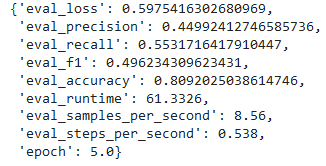
**3-3-1 نتایج آموزش مدل با 3200 جمله**

به جهت محدودیت منابع سخت افزاری[[24]](#footnote-24) شبکه به جای 30 ایپاک در 5 ایپاک آموزش داده شده است. نتایج ارزیابی حین آموزش به شرح جدول زیر می­باشد. همانطور که از نتایج زیر پیداست دقت مدل بعد از 5 ایپاک بالای 80% می باشد.

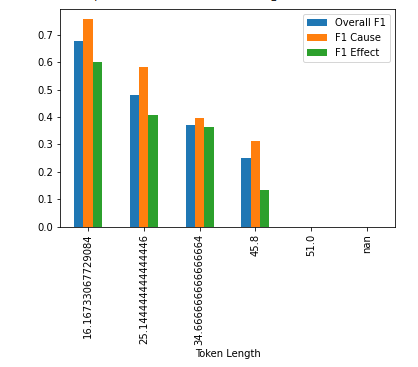


جدول ۲- خروجی معیارهای آموزشی مدل سوم

نتیجه ارزیابی روی داده های تست نیز به شرح پارامترهای ذیل می­باشد. همانطورکه اطلاعات زیر پیداست دقت مدل بر روی داده های تست بیش از 80% برآورد شده است.

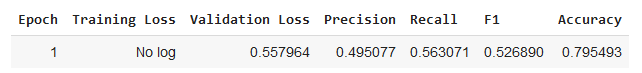


یکی از پارامترهای مهم بررسی طول توکن ها در دقت و کارآیی مدل می باشد. همانطور که از شکل زیر پیداست با کاهش طول توکن ها، دقت و کارایی مدل افزایش می­یابد.

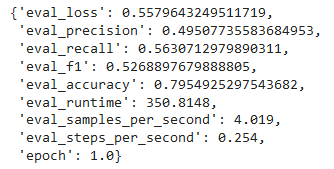


شکل 4- بررسی عملکرد مدل سوم تحت طول‌های مختلف توکن

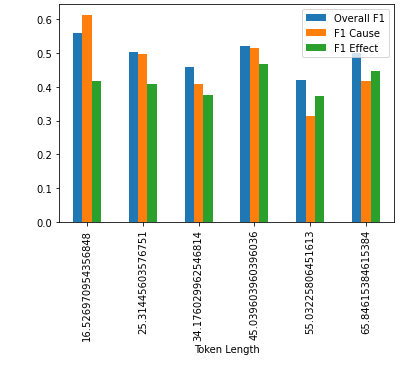
**3-3-2 نتایج آموزش مدل با 8800 جمله**

با عنایت به افزایش تعداد جملات به جهت محدودیت های سخت افزاری مدل یک ایپاک آموزش داده شده است و نتایج زیر حاصل شده است. مقایسه نتیجه ایپاک اول این مدل با مدل قبلی افزایش یک درصد در دقت را نشان می دهد.

نتیجه ارزیابی روی داده های تست نیز به شرح پارامترهای ذیل و دقت 79.5 درصد می­باشد.



شکل زیر دقت به تفکیک طول جملات نشان داده شده است.



شکل 5- بررسی عملکرد مدل سوم با 8800 جمله تحت طول‌های مختلف توکن

**3-3-3 جمع بندی مدل مبتنی بر BERT**

برای درک بهتر عملکرد مدل نمونه های متعددی را به عنوان جلمه ورودی برای تعیین روابط علت معلولی به مدل داده شده است که خروجی ذیل نشان از عملکرد قابل قبول مدل می­باشد. به نظر می رسد بکارگیری از یک مدل زبانی با قابلیتهای مکانیزم توجه و ایجاد بازنمایی مبتنی بر سیاق اثربخشی مدل و نتایج را بهبود داده است. نمونه خروجی مدل در فایل دمو به پیوست تحقیق اشاره شده است.ُ

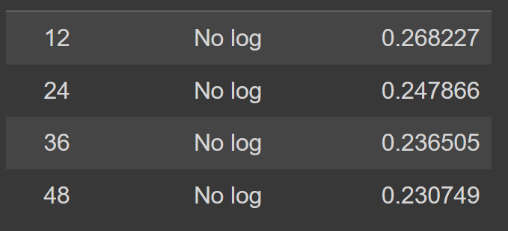
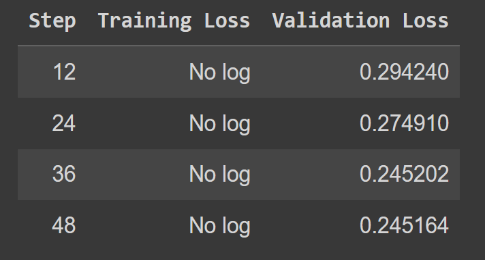
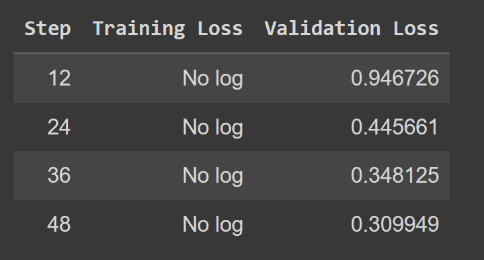
**۳-۴ شناسایی با استفاده از مساله‌ی پرسش‌وپاسخ استخراجی[[25]](#footnote-25)**

در این روش از مساله‌ی شناخته‌شده پرسش‌وپاسخ برای یادگیری و یافتن علت، معلول و نشانه‌گر استفاده شده‌است. به‌طور کلی در این مدل پرسش‌وپاسخ یک متن به ما ارایه شده و ما باید بر اساس آن پاسخ را بیابیم. توجه داریم که پاسخ نیز یک زیررشته از متن ورودی ما است. بدین مفهوم که خروجی از همان متن حاصل شده و استنتاجی نمی‌باشد. همچنین نحوه ارایه خروجی بر اساس یافتن توکن شروع و پایان پاسخ است. یادگیری آن نیز بدین صورت شکل می‌گیرد که به ازای هریک از متون یک سوال قرار داده و در ادامه آن به عنوان رشته ورودی به شبکه، متن (کانتکست) را قرار می‌دهیم. توکن‌های نماد، مانند فاصله‌دهنده را نیز میان آن‌ها قرار می‌دهیم. نحوه محاسبه خطا و امتیازدهی نیز بدین صورت شکل می‌گیرد که امتیاز یافتن صحیح توکن‌های شروع و پایان را بالا قرار می‌دهیم.

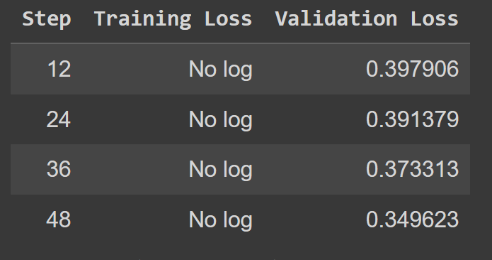
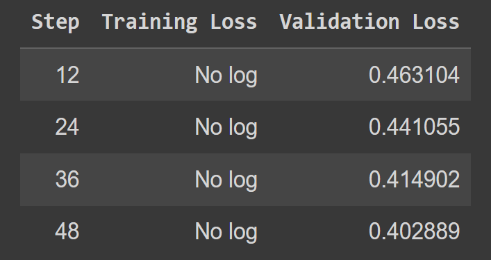
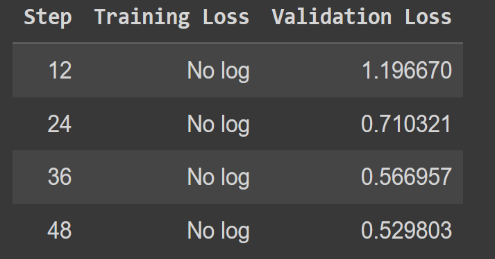
حال به‌ این موضوع می‌پردازیم که به چه نحوی از این مساله برای یادگیری و یافتن روابط علت-معلولی استفاده می‌کنیم. برای این مساله ۳ شبکه مجزای پرسش‌وپاسخ را برای هریک از علت، معلول و نشانه‌گر آموزش می‌دهیم. سوال مورد استفاده برای تمامی داده‌ها در شبکه اول ثابت بوده و رشته‌ای از کلمات ثابت، مانند علت، معلول، به‌دلیل و ... است. هرجمله یا عبارتی که می‌خواهیم از آن روابط علی را استخراج کنیم را نیز در ادامه آن قرار داده و به شبکه می‌دهیم. بدین صورت شبکه اول آموزش داده شده و نشانه‌گر از آن استخراج می‌شود.

برای شبکه دوم، نشانه‌گر یافت‌شده این‌بار به‌عنوان سوال قرارداده شده و متن در ادامه آن می‌آید. برای این شبکه هدف یافتن خروجی برای علت است. نهایتا همین کار مشابه را نیز برای شبکه سوم انجام داده و این ‌بار به طریق مشابه معلول را می‌یابیم. تنها ذکر چند نکته در این میان ضروری است. یک آن‌که در شبکه اول، برای یافتن نشانه‌گر، ممکن است حرف‌اضافه‌ای به‌عنوان نشانه‌گر وجود نداشته باشد، ازاین‌رو از فعل‌های علیت‌ساز و عبارات‌فعلی برای نشانه‌گذاری و یافتن رابطه علیت استفاده می‌کنیم، چراکه می‌دانیم که برخی از فعل‌ها نیازمند معلول و یا علت هستند[[26]](#footnote-26). همچنین نکته بسیار مهم دیگر آنکه نیاز به ارایه یک خروجی در هر مرحله نیز نمی‌باشد. می‌توان با تنظیم نحوه خروجی‌دادن براساس امتیاز، چندین خروجی و رابطه علیت را همزمان را درنظر گرفت. خروجی بر اساس ردکردن یک ترشهلد خاص از امتیاز یا دو مقدار بیشینه امتیاز و موارد مشابه دیگر از مثال‌های این مورد هستند.

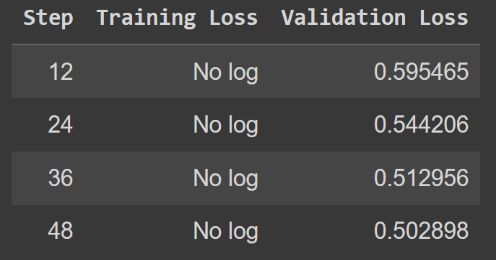
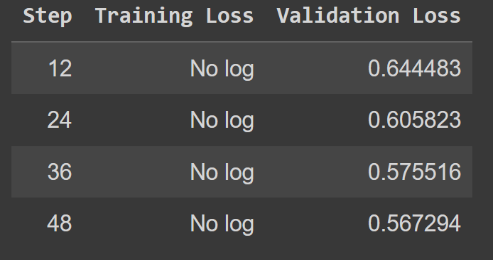
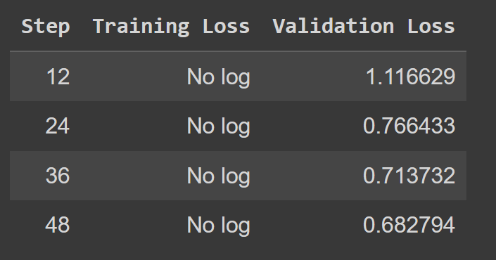
پیش از بررسی خروجی‌های مساله نگاهی دقیق‌تر به معماری و مدل مساله داشته باشیم. شبکه اولیه همان برت فارسی است که برای مساله پرسش‌وپاسخ تیون شده‌است. این مدل از hugging face استخراج شده و در ادامه کار به فاین‌-تیون‌کردن آن پرداخته‌ایم. برای این مدل و هریک از ۳ شبکه ۳۰۰۰ داده آموزش داشته که ابتدا آن‌ها را درهم میریزیم، سپس داده‌های آموزش را ۸۰۰ تا ۸۰۰ تا به شبکه داده و هربار مدل را برای بهینه‌بودن کار ذخیره و در ادامه بارگذاری می‌کنیم. Batch-size را نیز ۱۶ قرار می‌دهیم. ۶۰۰ داده نهایی را نیز به‌عنوان اعتبارسنج درنظرگرفته و برای تنظیم هایپرپارامترها به‌کارمی‌بریم. برای این کار نیز بعد از هر ۱۲ مرحله، اعتبارسنجی را انجام داده و نرخ آموزش را به مرور کاهش می‌دهیم. ۳۰۰ داده تست نیز استفاده شده که خروجی‌های آن را در ادامه می‌بینیم. خروجی‌ها برای هر شبکه و در ۳ مرحله آموزش ۸۰۰تایی آورده شده‌است. نهایتا همانطور که پیشتر نیز ذکر شد بر روی ۶۰۰ تای باقی‌مانده اعتبارسنجی انجام می‌شود.



شکل 6- خروجی آموزش ۳ مرحله‌ای شبکه اول (نشانه‌گر) مدل چهارم- نرخ یادگیری به‌ترتیب ۵،۵،۲.۵



شکل 7- خروجی آموزش ۳ مرحله‌ای شبکه دوم (علت) مدل چهارم- نرخ یادگیری به‌ترتیب ۵،۵،۵



شکل 8- خروجی آموزش ۳ مرحله‌ای شبکه سوم (معلول) مدل چهارم- نرخ یادگیری به‌ترتیب ۱۵،۵،۵

یکی از مزایای این روش این است که برچسب‌گذاری ثابت برای هر توکن نداریم. این مورد آن‌جا به کار می‌آید که یک عبارت در یک جمله ممکن است هم نقش علت به خود بگیرد و هم معلول. از‌این‌رو و باتوجه به نشانه‌گر مورد استفاده آن، نقش عبارت به دو حالت مجزای علت و معلول شکسته شده و توانایی یافتن هریک را باتوجه به نشانه‌گر مخصوص آن خواهیم داشت. به مثال‌ زیر توجه کنید:

همانطور که از مثال بالا برمی‌آید، چاقی دوبار، یک‌بار در قالب علت و بار بعد در قالب معلول ظاهر شده‌است. این نحوه از برچسب‌گذاری به‌دلیل تفاوت نشانه‌گرهای روابط علیت که یکی کلمه *ناشی* بوده و دیگری *باعث* می‌تواند به‌درستی هربار این نقش چاقی را به‌درستی تشخیص دهد.

|  |  |  |  |
| --- | --- | --- | --- |
| Seq matcher | Precision | Recall |  |
| ۹۱.۷ | ۹۶.۹ | ۹۰.۹ | **Cause** |
| ۹۱.۶ | ۹۵.۲ | ۸۸.۹ | **Effect** |
| ۹۶.۶ | ۹۳.۶ | ۹۲.۳ | **Marker** |

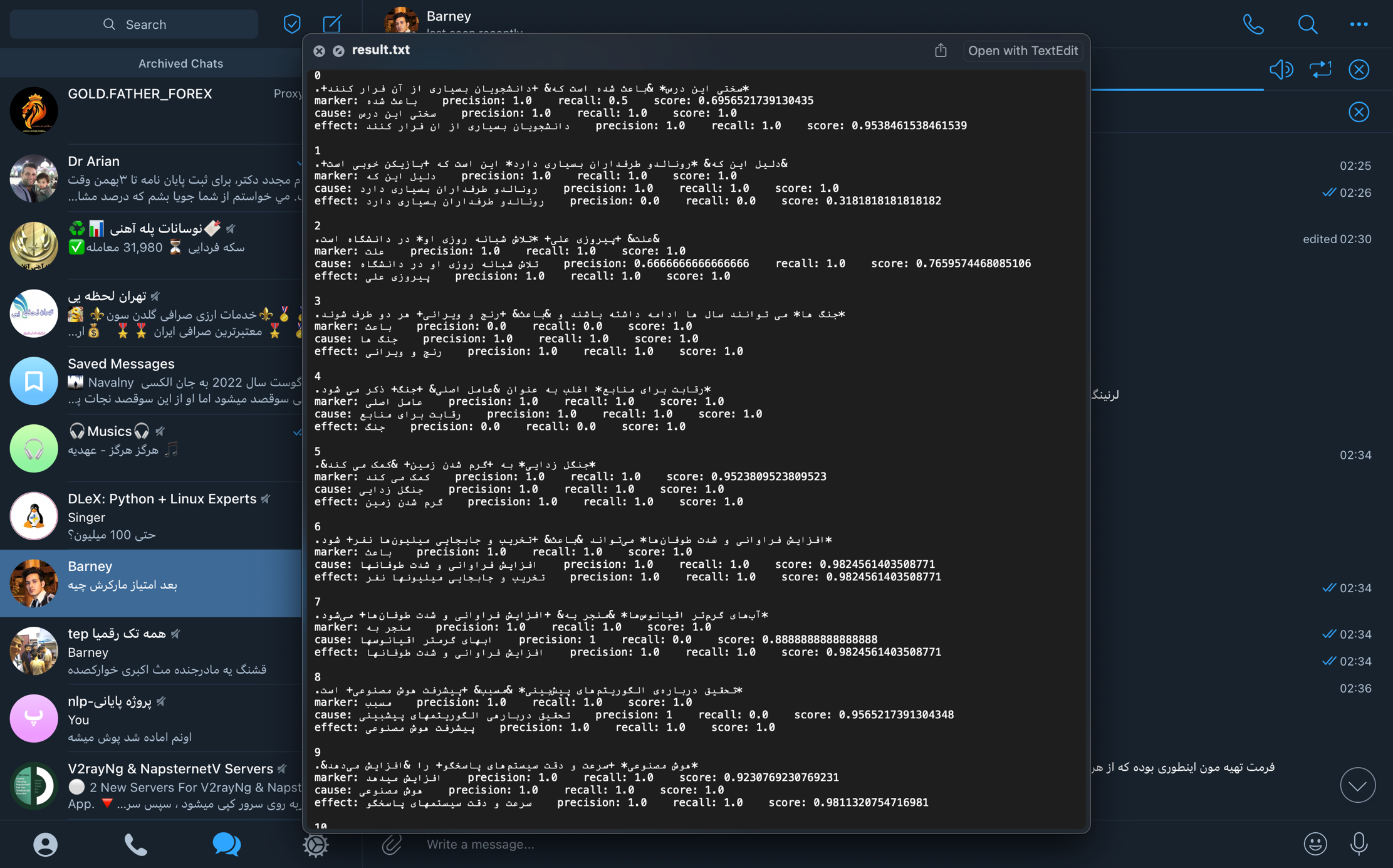
جدول ۳- عملکرد خروجی مدل چهارم برروی کل داده‌های تست

|  |  |  |  |
| --- | --- | --- | --- |
| Seq matcher | Precision | Recall |  |
| ۹۰.۹ | ۹۷.۴ | ۸۹.۸ | **Cause** |
| ۹۱.۱ | ۹۵.۷ | ۸۷.۸ | **Effect** |
| ۹۷.۳ | ۹۳.۹ | ۹۲.۳ | **Marker** |

جدول ۴- خروجی عملکرد مدل چهارم برروی داد‌ه‌های تست شامل علت و معلول

|  |  |  |  |
| --- | --- | --- | --- |
| Seq matcher | Precision | Recall |  |
| ۹۵.۵ | ۹۴.۸ | ۹۶ | **Cause** |
| ۹۳.۸ | ۹۲.۸ | ۹۴.۱ | **Effect** |
| ۹۳.۶ | ۹۲.۱ | ۹۲.۱ | **Marker** |

جدول ۵- خروجی عملکرد مدل چهارم برروی داده‌های تست بدون هیچ رابطه‌ علیتی



شکل 9- نمونه خروجی مدل چهارم

**۴- نتیجه‌گیری**

با توجه به اهمیت جایگاه مسئله استخراج راوبط علت و معلولی راهکارهای متعددی در حل این مسئله انتخاب شده است. تمرکز راهکار اول بر ایجاد بازنمایی کلمات ورودی و آموزش یک شبکه بازگشتی دو طرفه می باشد. در راهکار دوم از بازنمایی های موجود در زبان فارسی همانند FastText استفاده شده و مشابه روش اول با کمک یک شبکه بازگشتی در چارچوب یک مسئله بانظارت مسئله حل شده است. رویکرد سوم تمایز متفاوتی با دو رویکرد اول و دوم دارد. در این رویکرد مبتنی بر مدل های ترانسفورمر و قابلیتهایی که این مدل های با استفاده از بازنمایی های مبتنی بر سیاق و مکانیزم توجه ارائه می دهند مدل بر روی یک مدل پایه زبان فارسی آموزش داده شده است. نهایتا نیز رویکرد چهارم را داشتیم که براساس مدل پرسش‌وپاسخ بود که از یک دید دیگر به مساله نگریسته و خروجی‌هایی مبنی بر یافتن چند نقش یک عبارت ثابت بود.

سادگی و نیازی سخت افزاری حداقلی از جمله قابلیتهای روش های اول و دوم می باشد. هرچند دقت هر سه روش قابل قبول می باشد اما روش اول و سوم عملکرد نسبتا بهتری نسبت به روش دوم داشته اند. مدل سوم هم به جهت تعداد پارامترهای زیاد با چالش سخت افزار قوی تر چه در مرحله آموزش و چه در مرحله استنتاج مواجه هست. عملکرد مدل چهارم به تفکیک نشانگر، علت و معلول در گزارش قید شده است هرچند مانند مدل سوم همچنان نیاز به یکسری حداقل‌های بالای سخت‌افزاری را دارد.

از جمله اقدامات آتی می توان به موارد ذیل اشاره کرد:

* تنوع برچسب ها در داده ها
* تکمیل مجموعه داده های برچسب خورده با توجه به تجربیات بدست آمده
* اضافه کردن مکانیزم توجه به روش های اول و دوم جهت افزایش کارایی با توجه به مزیتهای این دو مدل از منظر سادگی و نیاز سخت افزاری کمتر
* تمیز‌سازی مناسب در مدل چهارم برای بهبود دقت

**مراجع**

[1] Abdu–Khader M. M. and Speight J. G., 2004, The concepts of energy, environment, and cost for process design, International Journal of Green Energy, vol. **1**, pp 137-151.

[2] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

[3] J Devlin, M Chang, K Lee and K Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 4171-4186, 2019.

[4] Li, Qian, Jianxin Li, Jiawei Sheng, Shiyao Cui, Jia Wu, Yiming Hei, Hao Peng et al. "A Compact Survey on Event Extraction: Approaches and Applications." *arXiv preprint arXiv:2107.02126* (2021).

[5] Zhou, Y., & Lei, J. (2019). Causal relation extraction with neural networks: A survey. arXiv preprint arXiv:1908.07835.

[6] Yang, J., Han, S.C. & Poon, J. A survey on extraction of causal relations from natural language text. *Knowl Inf Syst* **64**, 1161–1186 (2022). https://doi.org/10.1007/s10115-022-01665-w

[7] Drury, B., Gonçalo Oliveira, H., & De Andrade Lopes, A. (2022). A survey of the extraction and applications of causal relations. *Natural Language Engineering,* *28*(3), 361-400. doi:10.1017/S135132492100036X

[8] Vashishth, S., Karamanolakis, G., & Roth, D. (2019). Causal language and effect magnitude prediction in scientific text. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 1295-1305).

# [9] Chen, Q., Zhu, X., Ling, Z. H., Wei, S., Jiang, H., & Wang, T. (2020). A hybrid model based on neural networks for biomedical relation extraction. BMC Bioinformatics, 21(1), 1-11.

[10] ZeinabRahimi, MehrnoushShamsFard (2021). Persian Causality Corpus (PerCause) and the Causality Detection Benchmark. <https://arxiv.org/pdf/2106.14165.pdf>

[11] Khoo, Christopher SG, Syin Chan, and Yun Niu. (2000) .Extracting causal knowledge from a medical database using graphical patterns. Proceedings of the 38th annual meeting of the association for computational linguistics.

[12] Luo, Zhiyi, et al.(2016) .Commonsense causal reasoning between short texts.Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning.

1. Cause Effect Detection [↑](#footnote-ref-1)
2. Recurrent Neural Network [↑](#footnote-ref-2)
3. Transformer Model [↑](#footnote-ref-3)
4. Contextualized Representations [↑](#footnote-ref-4)
5. Self Attention [↑](#footnote-ref-5)
6. Deep Learning [↑](#footnote-ref-6)
7. Natural Language Processing [↑](#footnote-ref-7)
8. Relation Extraction [↑](#footnote-ref-8)
9. Convolutional Neural Network [↑](#footnote-ref-9)
10. End-to-End [↑](#footnote-ref-10)
11. Entities [↑](#footnote-ref-11)
12. Multi Task Learning [↑](#footnote-ref-12)
13. Transfer Learning [↑](#footnote-ref-13)
14. Pretrained Language Model [↑](#footnote-ref-14)
15. BIO [↑](#footnote-ref-15)
16. Supervised Learning [↑](#footnote-ref-16)
17. https://storage.googleapis.com/danielk-files/farsi-text/merged\_files/fawiki\_merged.txt [↑](#footnote-ref-17)
18. http://farsbase.net/download/PERLEX.zip [↑](#footnote-ref-18)
19. https://nlp.jhu.edu/causalbank/ [↑](#footnote-ref-19)
20. http://nlp.sbu.ac.ir/ [↑](#footnote-ref-20)
21. https://jipm.irandoc.ac.ir/article\_698606.html [↑](#footnote-ref-21)
22. Multi Head Attention [↑](#footnote-ref-22)
23. BertForTokenClassification [↑](#footnote-ref-23)
24. https://colab.research.google.com [↑](#footnote-ref-24)
25. Extractive Q&A [↑](#footnote-ref-25)
26. https://www.learngrammar.net/english-grammar/causative-verbs [↑](#footnote-ref-26)