Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
Volume 13, Number 4, Winter 2025, pp. 38-48
Received: 08/03/2024, Revised: 01/29/2025, Accepted: 02/24/2025

A Benchmark for Analyzing Knowledge Graph Embedding for Link Prediction Problem in Low-Resource Languages

Najmeh Torabian[1], Behrouz Minaei-Bidgoli[2*] and Mohsen Jahanshahi[3]
1- Department of Computer Engineering, Central Tehran Branch,
Islamic Azad University, Tehran, Iran.
2*- Department of Computer Engineering, Iran University of
Science and Technology, Tehran, Iran.
3- Department of Computer Engineering, Central Tehran Branch,
Islamic Azad University, Tehran, Iran.
[1]Najmeh.torabian@gmail.com, [2*] b_minaei@iust.ac.ir, and mjahanshahi@iauctb.ac.ir

Corresponding author's address: Behrouz Minaei-Bidgoli, Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran.

Abstract- Link prediction in knowledge graphs addresses predicting missing entities or relations of a knowledge graph, typically using knowledge graph embedding techniques. Training these models on low-resource-language knowledge graphs presents unique challenges, which have not been thoroughly addressed in the literature, and correspondingly there is no benchmark for evaluating link prediction methods on such graphs. These knowledge graphs often have unique topologies due to the characteristics of low-resource languages. Many knowledge graphs are derived from encyclopedias like Wikipedia, which in low-resource languages may have many propositions from common subjects and/or facts and few from less common ones, leading to distinctive topologies in the extracted knowledge graph. FarsBase, as the knowledge graph related to the Persian language, exemplifies these properties. Originating from Persian Wikipedia, it has some relation-types with many numbers of instances and some other relation types with very few instances. This paper introduces "FarsPishBin," a lightly-pruned version of FarsBase, as a benchmark for low-resource-language knowledge graph embedding. The authors argue that translational models are likely to outperform other embedding models on this benchmark. To check the mentioned hypothesis, the popular embedding models are evaluated on FarsPishBin and the experiments prove that translational models (as expected) perform best. This benchmark aims to serve as a standard platform for future-coming models addressing link prediction in low-resource-language knowledge graphs.

Keywords- Link prediction, Knowledge graph embedding, Link prediction benchmark

مجله علمی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
جلد ۱۳، شماره ۴، زمستان ۱۴۰۳ ، صفحه ۳۸-۴۸
دریافت: ۱۴۰۳/۰۵/۱۳، بازنگری: ۱۴۰۳/۱۱/۱۰، پذیرش: ۱۴۰۳/۱۲/۰۶

داده محکی برای تحلیل مدل‌های دگرنمایی گراف دانش برای مساله پیش‌بینی پیوند در زبان‌های کم‌منبع

نجمه ترابیان[۱]، بهروز مینایی-بیدگلی*[۲]، محسن جهانشاهی[۳]
۱- دانشکده مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد تهران مرکز، تهران، ایران.
*۲- دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران.
۳- دانشکده مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد تهران مرکز، تهران، ایران.
[1]najmeh.torabian@gmail.com, [2*]b_minaei@iust.ac.ir, [3]mjahanshahi@iauctb.ac.ir

* نشانی نویسنده مسئول: بهروز مینایی-بیدگلی نویسنده مسئول، تهران، رسالت، خیابان هنگام، خیابان دانشگاه ، دانشگاه علم وصنعت ایران، دانشکده مهندسی کامپیوتر.

چکیده- پیش‌بینی پیوند در بافتار گراف‌های دانش عبارت است از پیش‌بینی موجودیت‌ها یا روابط غیرموجود که معمولاً با استفاده از دگرنمایی گراف دانش است. آموزش این مدل‌ها بر روی گراف‌های دانش استخراج شده از زبان‌های کم‌منبع چالش‌های منحصربه‌فردی را ایجاد می‌کند که در پژوهش‌های پیشین به آن پرداخته نشده است و در نتیجه هیچ داده‌ی محکی نیز برای ارزیابی روش‌های پیش‌بینی پیوند در چنین گراف‌های دانشی وجود ندارد. گراف‌های دانش استخراج‌شده از زبان‌های کم‌منبع به دلیل ویژگی های این زبان‌ها، اغلب توپولوژی منحصربه‌فردی دارند. برای مثال، بسیاری از گراف‌های دانش از دایره‌المعارف‌هایی مانند ویکی‌پدیا استخراج شده‌اند و ویکی‌پدیای مربوط به زبان‌های کم‌منبع معمولا برای موجودیت‌های بسیاری توضیحات نامبسوط ارائه کرده باشد. این مسأله باعث می‌شود گراف‌های حاصل از آنها نیز در خصوص بسیاری از گره‌ها یال‌های اندکی داشته باشند و این مسأله توپولوژی متفاوتی از گراف‌های دانش معیار برای آنها ایجاد می‌کند. فارس بیس (تنها گراف دانش مرتبط با زبان فارسی)، نمونه‌ای از این گراف‌هاست. این مقاله نسخه‌ای از فارس‌بیس را تحت عنوان «فارس‌پیش‌بین»، به‌عنوان داده‌ی محکی برای دگرنمایی گراف دانش با زبان‌های کم‌منبع معرفی می‌کند. سپس، نویسندگان ادعا و استدلال می‌کنند که در گراف‌های دانش استخراج‌شده از زبان‌های کم‌منبع، مدل‌های فاصله انتقالی از سایر مدل‌های دگرنمایی گراف دانش بهتر عمل می‌کنند. برای آزمودن این فرضیه، مدل‌های دگرنمایی گراف دانش رایج بر روی آن آزمایش و ارزیابی شده و اثبات شد که مدل‌های فاصله انتقالی بهترین عملکرد را روی چنین داده‌ی محکی دارند. این داده‌ی محک انتظار می‌رود که زین‌پس به عنوان یک زیرساخت ارزیابی استاندارد برای پژوهش‌هایی که به پیش‌بینی پیوند در گراف‌های دانش با زبان‌های کم‌منبع خواهند پرداخت استفاده شود.

واژه‌های کلیدی: پیش‌بینی پیوند، دگرنمایی گراف دانش، داده‌ی محک پیش‌بینی پیوند

۱- مقدمه

گراف دانش، شکل ساختاریافته دانش به‌دست‌آمده از داده‌های مختلف غیرساخت‌یافته و نیمه‌ساخت‌یافته، از جمله متون دایره‌المعارفی و جداول است [۱]. گراف‌های دانش به دلیل کاربردهای گسترده و مزایای صنعتی قابل توجه در سال‌های اخیر بسیار مورد توجه قرار گرفته‌اند [۲].

۱-۱- دگرنمایی گراف دانش

گراف دانش یک پایگاه دانش است که هزاران یا میلیون‌ها موجودیت موجود در یک جهان معنایی را از طریق بازنمایی روابط و ویژگی‌هایشان به شکل گراف‌محور نمایش می‌دهد. کوچک‌ترین جزء دانشی در گراف دانش یال‌های گراف یا اصطلاحا سه‌تایی‌ها هستند. هر سه‌تایی شامل «موجودیت ابتدایی[۱]»، «رابطه‌ی برچسب‌دار دانشی» و «موجودیت انتهایی[۲]» است که موجودیت‌های ابتدایی و

انتهایی به عنوان گره‌ها و روابط برچسب‌دار به عنوان یال می‌باشد. به‌عنوان مثال گزاره‌ی دانشی «تهران پایتخت ایران است» را می‌توان با در نظر گرفتن «تهران» به‌عنوان موجودیت ابتدایی، «ایران» به‌عنوان موجودیت انتهایی، و رابطه‌ی «پایتخت بودن» به‌عنوان رابطه به‌صورت یک سه‌تایی نمایش داد. به‌عبارت دیگر، می‌توان گفت در گراف‌های دانش گزاره‌ها به صورت سه‌تایی‌های جهت‌دار برچسب‌دار نمایش داده می‌شود.

در زمینه گراف دانش، مفهوم پیش‌بینی پیوند به فرآیند کامل کردن اطلاعات ناموجود یک گراف دانش، از جمله موجودیت‌ها یا روابط غایب می‌پردازد. این کار معمولاً بر دگرنمایی گراف دانش (KGE[3]) متکی است.

دگرنمایی گراف دانش به فرآیندی اطلاق می‌شود که در آن گره‌ها و یال‌های یک گراف دانش به بردارهای عددی در فضای برداری تبدیل می‌شوند. این بردارها به گونه‌ای طراحی شده‌اند که اطلاعات ساختاری و معنایی گراف را حفظ کنند.

یکی از مزایای اصلی دگرنمایی گراف دانش این است که با تبدیل گراف دانش به فضای برداری، می‌توان از الگوریتم‌های یادگیری ماشین و شبکه‌های عصبی برای تحلیل و پردازش داده‌های گراف استفاده کرد. مدل‌های دگرنمایی گراف دانش روابط پیچیده بین موجودیت‌ها را بهتر درک نموده، پیش‌بینی‌های دقیق‌تری انجام می‌دهند.

روش‌های مختلفی برای دگرنمایی گراف دانش وجود دارد که هر کدام مزایا و معایب خاص خود را دارند. مدل‌های KGE خود را با TransE معرفی می‌کنند که در دسته‌ای از مدل‌ها تحت عنوان «مدل‌های فاصله انتقالی[۴]» قرار می‌گیرد [۳].

پس از TransE، بسیاری از مدل‌های فاصله انتقالی پیشنهاد شدند که قصد داشتند به تدریج مسیر آغاز شده را توسط مدل TransE، از طریق تبدیل سه‌تایی‌ها به بردار و پیش‌بینی سه‌تایی‌های جدید در گراف‌های دانش با در نظر گرفتن مکان بردارهای دگرنمایی‌شده، تکمیل کنند.

به موازات مدل‌های فاصله انتقالی، گروهی دیگر از مدل‌ها تحت عنوان «مدل‌های تطبیق معنایی[۵]» ارائه شد که تکنیک‌های تجزیه ماتریس را برای دگرنمایی مؤلفه‌های سه گانه مجموعه آزمون و پیش‌بینی سه‌تایی‌های جدید اعمال می‌کنند.

دسته پرکاربرد سومی نیز از مدل‌های پیش‌بینی پیوند وجود دارد که به آن «مدل‌های یادگیری عمیق[۶]» اطلاق می‌شود. این مدل‌ها که از آن تحت عنوان «مدل‌های شبکه عصبی عمیق» نیز یاد می‌شود، رفتار سه‌تایی‌ها را در یک گراف دانش از طریق یادگیری عمیق، یاد گرفته، سه‌تایی‌های جدیدی را پیشنهاد نموده، و بر اساس آنها گراف دانش را تکمیل می‌کنند.

به‌صورت عمومی، با وجود اینکه مدل‌های شبکه عصبی عمیق از دو دسته قبلی پرهزینه‌تر و زمان‌برتر هستند، در گراف‌های دانش استاندارد عملکرد پیش‌بینی قوی‌تری از خود نشان می‌دهند.

۱-۲- گراف‌های دانش کم‌منبع

چالش اصلی با گراف‌های دانش در زبان‌های کم‌منبع در توپولوژی متفاوت آنها در مقایسه با گراف‌های دانش استاندارد نهفته است. این گراف‌ها اغلب از ویکی‌پدیا تولید می‌شوند. در مورد زبان‌های کم‌منبع، این دایره المعارف‌ها ویژگی‌های متمایزی از خود نشان می‌دهند.

به عنوان مثال، آنها ممکن است سه‌تایی‌های متعددی را برای روابط پرکاربرد در بر داشته باشند و در عین حال برای روابط کم‌استفاده مصادیق بسیار کمی در بر داشته باشند. این ویژگی‌های خاص منجر به توپولوژی‌های متمایز در گراف‌های دانش کم‌منبع می‌شود و آنها را از گراف‌های دانش استاندارد جدا می‌کند.

آموزش مدل‌های KGE بر روی گراف‌های دانش کم‌منبع، چالش‌های منحصربه‌فردی را برای پیش‌بینی پیوند ارائه می‌کند که به اندازه کافی در پژوهش‌های پیشین به آن پرداخته نشده است. لذا، درحال‌حاضر هیچ داده‌ی محکی[۷] برای ارزیابی روش‌های پیش‌بینی پیوند در چنین گراف‌های دانشی وجود ندارد.

۱-۳- تبیین مسأله

زبان فارسی یک زبان کم‌منبع است و تنها، دو پایگاه واژگانی/دانشی تا کنون بر روی این زبان توسعه داده شده است؛ یکی از آنها پایگاه واژگانی فارس‌نت [۴] و دیگری گراف دانش فارس‌بیس[۸] [۵] می‌باشد که اولی تنها پایگاه واژگانی و دومی تنها گراف دانش موجود برای این زبان می‌باشد.

با توجه به اینکه گراف دانش فارس‌بیس از ویکی‌پدیای فارسی ساخته شده و ویکی‌پدیای فارسی چند برابر کوچکتر از ویکی‌پدیای زبان‌های پرمنبعی همچون انگلیسی است، برخی از روابط فارس‌بیس دارای مصادیق بسیار کم و برخی دیگر از آنها دارای مصادیق بسیار زیاد هستند و این توپولوژی خاص، مسأله پیش‌بینی پیوند را با چالش مواجه می‌نماید.

در این مقاله، با تشخیص نیاز به داده‌ی محکی در پیش‌بینی پیوند

برای گراف‌های دانش کم‌منبع، یک نسخه کوچک‌شده از فارس‌بیس را با نام «فارس‌پیش‌بین» معرفی می‌کنیم. برای این منظور، مجموعه‌ای از مراحل را برای ایجاد تغییرات در فارس‌بیس برای ایجاد یک داده‌ی محک پیش‌بینی پیوند ارائه می‌کنیم. این مجموعه‌داده با تمرکز بر دگرنمایی گراف دانش در زبان‌های کم‌منبع طراحی شده است.

در ادامه، فرضیه‌ای مطرح می‌شود که ادعا می‌کند برای پیش‌بینی پیوند مبتنی‌بر دگرنمایی گراف دانش، اگر گراف دانش مورد بررسی کم‌منبع باشد، روش‌های ساده‌تر و دارای پیچیدگی کمتر بهتر عمل می‌کنند؛ که این فرضیه در پایان مقاله اثبات می‌گردد.

۱-۴- ساختار مقاله

ساختار این مقاله به شرح زیر است: بخش دوم پژوهش‌های مرتبط را مورد بحث قرار می‌دهد. بخش سوم ساختار داده‌ی محک پیش‌بینی پیوند فارسی پیشنهادی و فرضیه نویسندگان در خصوص رفتار مدل‌های KGE در داده‌ی محک پیشنهادی را شرح می‌دهد.

بخش چهارم آزمایش‌های انجام‌شده بر روی فارس‌پیش‌بین را ارائه نموده، سپس، نتایج به‌دست‌آمده برای تحلیل و بررسی فرضیه ارائه‌شده در بخش ۳-۳ و سایر نکات قابل استنتاج از آزمایش‌ها را توضیح می‌دهد. در نهایت، این مقاله در بخش پنجم با نتیجه‌گیری از مقاله به‌همراه بحث در مورد تحقیقات آینده به پایان می‌رسد.

۲- بررسی ادبیات موضوع

از آنجا که گراف‌های دانش برای درک و استدلال با زبان‌های ماشینی ایجاد می‌شوند، دانش بشری باید به شکلی قابل پردازش برای ماشین نمایش یافته، ذخیره و استخراج شود. گراف‌های دانش به عنوان پایگاه‌های دانش برای موجودیت‌ها و روابط بین این موجودیت‌ها ایجاد شده است. با این حال، گراف‌های دانش در مورد بیشتر زبان‌های انسانی محدودیت‌هایی دارند که حاصل از کم‌منبع بودن آن زبان‌ها می‌باشند.

علاوه بر این، دگرنمایی گراف دانش در طول سال‌ها پیشرفت خارق‌العاده‌ای در بازنمایی دانش ارائه کرده است، اما در بسیاری از کاربردهای موجود در دنیای واقعی، مدل‌های KGE هنوز نیاز به ارتقا دارند. اندازه، پیچیدگی و تنوع این مجموعه داده‌ها به طور قابل توجهی استفاده از آنها را به چالش می‌کشد.

۲-۱- گراف‌های دانش شناخته‌شده

در این بخش، گراف‌های دانش از دیدگاه زبانی به اختصار مورد بررسی قرار می‌گیرند. اگرچه کارهای زیادی برای ایجاد گراف‌های دانش غیرانگلیسی یا چند زبانه انجام شده است، اما هیچ داده‌ی محکی برای پیش‌بینی پیوند در زبان‌های کم‌منبع یافت نشده است. به همین دلیل در این زیربخش ابتدا به گراف‌های دانش بین‌المللی دارای منابع زبانی غنی و سپس به گراف‌های دانش خاص‌زبانه استاندارد (غیر مربوط به زبان‌های کم‌منبع) می‌پردازیم.

۲-۱-۱- گراف‌های دانش بین‌المللی

گراف‌های دانش متعددی برای زبان‌های پرمنبع توسعه‌یافته است. با این حال، در بین زبان‌های پرمنبع این زبان انگلیسی است که حتی در گراف‌های دانش شناخته‌شده‌ای از جمله DBpedia [۶]، FreeBase [۷]، BabelNet [۸]، Wikidata [۹]، YAGO4 [۱۰]، Google Knowledge Graph [۱۱] به‌طور واضحی فراوانی بیشتری را دارا می‌باشد.

با این حال، چندین تلاش برای ساختن گراف‌های دانش به زبان‌های دیگر انجام شده است که نمونه‌هایی از آن در قسمت بعدی آورده شده است.

نکته حائز اهمیت در مورد این گراف‌های دانش این است که هیچ گزارشی مبنی بر استفاده از این موارد در دگرنمایی گراف دانش و پیش‌بینی پیوند وجود ندارد. به‌عبارت دیگر، هیچ یک از گراف‌های دانش ارائه شده در قسمت پیش رو، به‌عنوان داده‌ی محکی برای KGE ارائه نشده‌اند.

۲-۱-۲- گراف‌های دانش خاص زبان‌های مختلف

گراف دانش XLore یک گراف دانش دوزبانه انگلیسی-چینی است که دانش چینی را با استفاده از ویکی‌پدیای آن زبان غنی می‌کند [۱۲-۱۱].

یکی دیگر از گراف‌های دانش غیر انگلیسی Lynx است که گراف دانشی حقوقی در گراف دانش اروپایی چندزبانه و بخشی از یک پروژه بزرگ برای خدمات انطباق نوآورانه است [۱۳].

علاوه بر این، HOLINET یک گراف دانش کل‌نگر برای زبان فرانسه است [۱۴]. همچنین PolarisX یک گراف دانش بسط‌یافته خودکار[۹] است [۱۵]. ضمنا، RezoJDM [۱۶] معرفی به زبان فرانسوی و VisualSem [۱۷] گراف دانشی بصری است که مفهوم چندزبانی را از گستره‌ی زبان‌های طبیعی به نوع انتقال مطلب (متنی/دیداری) گذار داده است.

۲-۲- مدل‌های دگرنمایی گراف دانش در پیش‌بینی پیوند

مدل‌های KGE راهی برای دگرنمایی یک گراف دانش در فضای برداری با حفظ ویژگی‌های اصلی آن هستند که یکی از مهم‌ترین کاربردهای آن پیش‌بینی پیوند می‌باشد.

این مدل‌ها به سه دسته اصلی تقسیم می‌شوند که در ادامه به اختصار به آنها پرداخته می‌شود.

۲-۲-۱- مدل‌های فاصله انتقالی

مدل‌های فاصله انتقالی اولین گروه از مدل‌های KGE هستند. این مدل‌ها همانند سایر مدل‌های دگرنمایی دانش وظیفه‌ی دگرنمایی سه‌تایی‌های دانشی را برعهده دارند. سه‌تایی‌های دانشی عبارتند از موجودیت ابتدایی، رابطه، و موجودیت انتهایی. مدل‌های فاصله‌ی انتقالی هر یک از سه عنصر «موجودیت ابتدا»، «رابطه» و «موجودیت انتهایی» را به عنوان یک بردار در فضا دگرنمایی می‌کنند. سپس با توجه به مدل خاص انتخاب شده پیش‌بینی‌ها انجام می‌شوند.

اولین مدل در این دسته TransE است که موفقیت چشم‌گیری کسب کرده است [۳]. این مدل مسأله دگرنمایی موجودیت‌ها و روابط داده‌های چند رابطه‌ای در فضای برداری با ابعاد کم را مد نظر قرار داده است که به راحتی قابل آموزش می‌باشد.

با توجه به تعداد پارامتر کم مورد استفاده، این روش بر روی گراف‌های دانش بزرگ قابل اجرا می‌باشد. سعی بر آن دارد که موجودیت‌های ابتدا و انتهای یک سه‌تایی را به‌گونه‌ای دگرنمایی نماید که در فضای جدید نزدیک هم باشند، یعنی از مجموع موجودیت ابتدا و رابطه بتوان موجودیت انتها را به دست آورد یا به عبارت دیگر رابطه $h + r \approx t$ برقرار باشد که در آن h نمایانگر موجودیت ابتدا، r نمایانگر رابطه و t نمایانگر موجودیت انتها می‌باشد.

علی‌رغم سادگی و تعداد پارامتر کم، این الگوریتم مشکلاتی از جمله در بر نگرفتن روابط یک به چند (1-to-N)، چندبه‌یک (N-to-1) و چند به چند (N-to-N) می‌باشد.

برای پوشش دادن مشکل فوق، مدل‌های دیگر فاصله انتقالی از جمله TransD [۱۸]، TransH [۱۹]، TransR [۱۸] و RotatE [۲۰] ارائه شدند که یکی پس از دیگری، سعی در بهبود کارایی مدل‌های پیشین خود دارند.

۲-۲-۲- مدل‌های تطبیق معنایی

مدل‌های تطبیق معنایی بر مدل‌سازی شباهت معنایی بر اساس روابط بین موجودیت‌ها استوار می‌باشند. اولین مدل در این مجموعه RESCAL است که در آن موجودیت‌ها به عنوان بردار و روابط بین این موجودیت‌ها ماتریسی در نظر گرفته شده، بر مبنای فاکتورگیری بر روی تنسورهای[۱۰] سه بعدی عمل می‌کند [۲۰]. در این مدل هر موجودیت با یک بردار برای نگهداری فاکتور پنهان[۱۱] و هر رابطه با یک ماتریس تعامل[۱۲] میان فاکتورهای پنهان بازنمایی می‌شوند. در این مدل، تابع امتیازدهی توسط تابع دوجهته زیر تعریف می‌شود:

$$f_r(h.t) = h^T M_r t = \sum_{i=0}^{d-1}\sum_{j=0}^{d-1}[M_r]_{ij}.[h]_i.[t]_j \qquad (۱)$$

که در آن h بیانگر بردار مربوط به موجودیت ابتدا، t بردار مربوط به موجودیت انتها، و M ماتریس مربوط به روابط می‌باشد. این تابع امتیازدهی به ازای هر دو موجودیت انتخابی محاسبه می‌شود. به‌این‌صورت از روی شباهت معنایی سه‌تایی‌های جدید پیشنهاد می‌شود و موجودیتی که امتیاز بالاتری از نظر شباهت داشته باشد، پیشنهاد پیوند جدید خواهد بود.

مدل‌های دیگر این دسته که سعی در بهبود عملکرد مدل‌های پیش از خود دارند عبارتند از DistMult [۲۱]، HolE [۲۲] و ComplEx [۲۳].

در این مدل‌ها، داده‌های رابطه‌ای به‌عنوان یک ماتریس سه‌بعدی X با اندازه $n \times n \times m$ مدل‌سازی می‌شوند، که n تعداد موجودیت‌ها و m تعداد روابط است.

۲-۲-۳- مدل‌های شبکه عصبی

مدل‌های شبکه عصبی از ابزارهای شبکه عصبی عمیق برای یادگیری مدل سه تایی در گراف دانش استفاده کرده، سه‌تایی‌های جدید را پیشنهاد نموده، گراف دانش را تکمیل می‌کنند. یکی از اولین روش‌ها برای دگرنمایی گراف دانش مبتنی بر یادگیری عمیق، مدل ProjE است که در آن موجودیت‌ها و روابط بین آنها به طور یکپارچه دگرنمایی شده است [۲۴].

از این مدل در عمل بخاطر داشتن مشکل سرعت روش و دیر همگرا شدن استقبال نشد اما پس از آن مدل‌های شبکه عصبی دیگری از جمله ConvKB [۲۵]، ConvE [۲۶] ارائه شدند. تابع امتیازدهی مدل ConvE به شکل زیر تعریف شد.

$$\psi(e_s.e_o) = f(V^T.W).e_o \qquad (۲)$$

که در آن e_s و e_o بردارهای دگرنمایی شده‌ی موجودیت ابتدا و انتها

است؛ V بردار اتصال‌یافته‌ی بین موجودیت ابتدا و رابطه‌ی بین دو موجودیت می‌باشد که از طریق روش‌های پیچشی ایجاد شده است؛ و W نیز ماتریس وزن‌های مدل می‌باشد.

در ادامه، مدل‌هایی برای گسترش ایده‌ی مربوط به ConvE از جمله SACN [۲۷] توسعه داده شدند. مدل‌های شبکه عصبی عمیق اگرچه در مقایسه با سایر دسته‌ها زمان‌برتر هستند، عملکرد خوبی را در فرآیند پیش‌بینی از خود نشان داده‌اند.

۳- ساخت داده‌ی محک

همانطور که در بخش مقدمه ذکر شد، حل مساله پیش‌بینی پیوند عمدتاً از طریق دگرنمایی گراف دانش انجام می‌شود. مسأله‌ی پرداخته‌شده در این مقاله، مسأله پیش‌بینی پیوند در گراف‌های دانش کم‌منبع است.

این گراف‌های دانش ویژگی‌های خاصی دارند که منجر به مشکلات ویژه در فرآیند دگرنمایی گراف دانش می‌شود. با توجه به اینکه این مشکلات در پژوهش‌های پیشین بررسی نشده‌اند، در این بخش، ما یک داده‌ی محک برای KGE با زبان کم‌منبع پیشنهاد می‌کنیم.

با فرض اینکه ویکی‌پدیای فارسی به‌عنوان نمونه‌ای واقعی برای زبان‌هایی با منابع محدود (ویکی‌پدیای فارسی یک میلیون مقاله دارد که شش برابر کمتر از نسخه انگلیسی است)، گراف دانش ایجاد شده از ویکی‌پدیای فارسی نمونه مناسبی از منابع کم خواهد بود.

فارس‌بیس که تنها گراف دانش ساخته شده از ویکی‌پدیای فارسی است به عنوان یک گراف دانش با زبان کم‌منبع شناخته می‌شود. بنابراین، خارج از انتظار نیست که برخی از روابط کم‌کاربرد در این گراف دانش، فقط چند مصداق دارند و برخی از روابط پرکاربرد، نمونه‌های متعددی را در این گراف دانش دارا هستند.

باتوجه به مقدمه ذکرشده، فارس‌بیس نامزد خوبی برای داده‌ی محک KGE برای زبان‌های کم‌منبع به نظر می‌رسد. با این حال، مانند بسیاری دیگر از گراف‌های دانش کم‌منبع، فارس‌بیس دو اشکال اصلی برای استفاده از مدل‌های KGE دارد:

(۱) فراداده‌های میلیونی‌ای که برای فرآیند دگرنمایی گراف دانش اضافی هستند و باید قبل از تبدیل به یک داده‌ی محک پیش‌بینی پیوند حذف شوند.

(۲) بخش قابل توجهی از روابط دارای مصادیق ناکافی برای پردازش از طریق مدل‌های KGE هستند (٪۷۱ از آنها کمتر از ۱۰۰ مصداق دارند، که در میان آنها بسیاری از آنها فقط به یک مصداق در هر نوع رابطه محدود می‌شود).

به همین ترتیب، در این بخش، یک نوع هرس‌شده از فارس‌بیس به نام فارس‌پیش‌بین را معرفی می‌کنیم. برای این منظور، مجموعه‌ای از اقدامات را برای ارتقاء فارس‌بیس به یک داده‌ی محک پیش‌بینی پیوند با زبان کم‌منبع، تشریح می‌کنیم.

۳-۱- ساخت فارس‌پیش‌بین

در این بخش، فرآیند تبدیل گراف دانش فارس‌بیس به داده‌ی محک پیش‌بینی پیوند فارس‌پیش‌بین را توضیح می‌دهیم. این فرآیند را می‌توان در چهار مرحله کلی خلاصه کرد که روند کلی آن در شکل ۱ مشخص شده است.

(مرحله A) هرس‌کردن روابط غیر سه‌تایی؛
(مرحله B) حذف گراف‌های فرعی جنگل؛
(مرحله C) حذف روابط کم‌کاربرد؛
(مرحله D) حذف موجودیت‌های کم‌کاربرد.

مرحله اول شامل حذف سه‌تایی است، به جز آنهایی که دارای دو موجودیت (در قسمت موجودیت ابتدایی و انتهایی) هستند.

(الف) «فراداده‌ها» به سه‌تایی‌هایی از گراف اطلاق می‌شود که یک طرف آن یک موجودیت و طرف دیگر آن یک پیوند[۱۳] به یک مفهوم بیرونی است که برای توضیح یا توصیف بیشتر آن موجودیت در گراف‌های دانش قرار می‌گیرد.

(ب) «روابط هستان‌شناسانه[۱۴]» به سه‌تایی‌هایی از گراف گفته می‌شود که یک طرف آن یک موجودیت و طرف دیگر آن یک کلاس از گراف دانش است. مثلا یک بازیکن فوتبال ممکن است یک یال به کلاس «ورزشکار» در گراف دانش داشته باشد.

(ج) «موجودیت‌های غیر فارسی» موجودیت‌هایی هستند که تنها نام لاتین دارند و خارج از فرهنگ و زبان گراف دانش فارسی به این گراف دانش اضافه شده‌اند.

(د) «ویژگی‌های[۱۵] موجودیت‌ها» سه‌تایی‌هایی هستند که یک طرف آن‌ها یک موجودیت و طرف دیگر آن‌ها یک مقدار[۱۶] رشته‌ای یا عددی است که توصیفی از آن موجودیت ارائه می‌دهد.

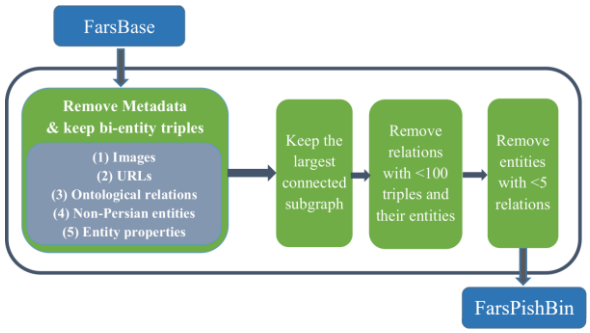
وجه مشترک چهار نوع سه‌تایی فوق در این است که همه‌ی آن‌ها سه‌تایی‌های تک‌موجودیتی هستند و بیان‌گر ارتباط بین هیچ دو موجودیتی نیستند. با توجه به این مسأله و همچنین این حقیقت که روش‌های دگرنمایی گراف دانش مبتنی‌بر یادگیری روابط بین جفت‌موجودیت‌ها عمل می‌کنند، چهار نوع سه‌تایی فوق را از گراف دانش حذف می‌نماییم.

پس از اجرای این چهار عنوان، تعداد سه‌تایی‌ها از ۵.۲ میلیون به

۵.۱ میلیون کاهش می‌یابد.

مرحله دوم به اصلاح ساختار گراف اشاره دارد. همان‌طور که اشاره شد، در مرحله‌ی اول حدود یک میلیون یال از گراف دانش حذف شدند. طبیعتا این مسأله با در نظر گرفتن توپولوژی خاص گراف‌های دانش زبان‌های کم‌منبع باعث خلوت شدن گراف و در نتیجه ایجاد مجموعه‌ای از زیرگراف‌های جدا از هم می‌شود.

اکثر این زیرگراف‌ها دارای گره‌های بسیار کمی هستند. با نگاهی به زیرگراف‌ها، مشاهده می‌شود که در بین ۴۸ زیرگراف مذکور، بزرگترین آن‌ها ۹۹.۸۷٪ از کل گراف دانش را شکل داده، ۴۷ جزء دیگر آن کمتر از یک درصد از سه‌تایی‌های گراف دانش را پوشش می‌دهند. به همین دلیل، در این مرحله تنها بزرگ‌ترین زیرگراف مذکور نگاه داشته شده و سایر زیرگراف‌ها حذف می‌گردند.



شکل ۱: نمودار بلوکی مربوط به فرآیند تبدیل گراف‌دانش فارس‌بیس به معیار پیش‌بینی پیوند فارس‌پیش‌بین

مرحله سوم روابط کم‌تعداد در گراف دانش را هرس می‌نماید چرا که گراف به‌دست‌آمده از مرحله قبل علاوه بر نامنظم بودن، نشان می‌دهد که تعداد سه‌تایی‌های موجود در رابطه‌های مختلف دارای تنوع زیادی است. با این حال، با توجه به اینکه بسیاری از روابط که به ندرت ظاهر می‌شوند نمی‌توانند اطلاعات مفیدی را در فرآیند دگرنمایی گراف دانش ارائه دهند.

با نگاهی به آماره‌های مربوط به گراف حاصل از مرحله قبل مشاهده می‌شود که حدود ۵٪ از روابط (۸۳ رابطه) حدود ۹۰٪ از سه‌تایی‌ها را شامل می‌شوند و ۹۵٪ دیگر (۱۷۰۶ رابطه) صرفا حدود ۱۰٪ از سه‌تایی‌ها را شامل می‌گردند. لذا، در این مرحله، ۹۵٪ مذکور که معادلا روابط با کمتر از ۱۰۰ مصداق هستند حذف می‌گردند.

مرحله چهارم به حذف زنجیره‌های موجودیت-موجودیت اختصاص دارد. توضیح آنکه بیش از ۹۰٪ موجودیت‌های گراف دانش فارس‌بیس دارای درجه‌ی کمتر یا مساوی با پنج می‌باشند که باعث ایجاد تعداد زنجیره‌های گرافی ناهمبند می‌شود. بنابراین این نوع از گره‌ها برای مسأله همگنی چالش‌آفرین هستند.

لذا، در این مرحله، موجودیت‌هایی که به کمتر از پنج موجودیت دیگر مرتبط هستند، برای جلوگیری از زنجیره‌های ذکر شده از گراف دانش حذف می‌شوند.

در نهایت پس از طی مراحل استانداردسازی پیش‌گفته، داده محک جدیدی (به نام فارس‌پیش‌بین) ایجاد می‌شود. باتوجه به حذفیات انجام‌شده در مراحل گذشته، فرم کلّی داده‌ی محک فارس‌پیش‌بین عملا فهرستی از سه‌تایی‌های دانشی است که دو طرف همه‌ی سه‌تایی‌ها موجودیت می‌باشد.

به‌منظور محک روش‌های پیش‌بینی پیوند مبتنی‌بر دگرنمایی دانش، بخشی از داده محک فارس‌پیش‌بین به‌عنوان داده آموزش و بقیه آن به‌عنوان داده آزمون در نظر گرفته شده، مدل‌های دگرنمایی گراف دانش روی قسمت اول آموزش یافته، و روی قسمت دوم با هم مقایسه می‌گردند.

۳-۲- آماره‌های فارس‌پیش‌بین

در این بخش آماری در مورد موجودیت‌ها و روابط مجموعه‌داده حاصل (فارس‌پیش‌بین) ارائه شده و مشخصات آن مورد بحث قرار می‌گیرد. پیش از آن‌که در قالب یک جدول، آماره‌های مذکور را بیان کنیم، ذکر این نکته شایسته است که پایگاه داده‌ی حاصل‌شده یک گراف دانش وابسته به فارس‌بیس است؛ و بنابراین هستان‌شناسی مستقل خود را ندارد و عملا همان ۷۶۷ کلاس فارس‌بیس را به ارث می‌برد.

جدول ۱: تعداد انواع رابطه، موجودیت‌ها، سه‌تایی‌ها و نسبت سه‌تایی‌ها به موجودیت و نسبت سه‌تایی‌ها به رابطه، مربوط به گراف‌های دانش فارس‌بیس، فارس‌پیش‌بین و فری‌بیس

داده محک	موجودیت	رابطه	سه‌تایی	سه‌تایی /موجودیت
فارس‌بیس	754,977	7,378	2,398,999	3
فارس‌پیش‌بین	107,827	392	622,287	6
FB15K	14,951	1,345	592,213	40

جدول ۱ تعداد رابطه‌ها، موجودیت‌ها و سه‌تایی‌های فارس‌بیس، فارس‌پیش‌بین و فری‌بیس[۱۷] و نسبت سه‌تایی به موجودیت و سه‌تایی به روابط در هر یک از آنها را ارائه می‌کند.

با توجه به اطلاعات ارائه شده در جدول ۱، در حالی که گراف دانش فارس‌بیس به طور متوسط برای هر موجودیت سه عدد سه‌تایی دارد،

نسبت مشابه در داده‌ی محک فارس‌پیش‌بین شش می‌باشد.

اگرچه این واقعیت نشان می‌دهد که فارس‌پیش‌بین تقریباً دو برابر بیشتر از فارس‌بیس همگن است، اما مقایسه این نسبت با نسبت مشابه در یک داده‌ی محک پیش‌بینی پیوند استاندارد مانند فری‌بیس (دارای نسبت ۴۰) نشان می‌دهد که این داده‌ی محک پیش بینی پیوند با زبان کم‌منبع، بسیار پراکنده است.

۳-۳- فرضیه

در این بخش، فرضیه‌ای در مورد نحوه برخورد با مدل‌های KGE بر روی داده‌ی محک پیش‌بینی پیوند فارس‌پیش‌بین ارائه شده است. در گراف‌های دانش خلوت (تنک) که خلوتی آنها به‌خاطر کم‌منبع بودن زبان است، (برخلاف گراف‌های استخراج‌شده از زبان‌های با منابع غنی) ناموجود بودن روابط در گراف به‌معنای برقرارنبودن آن روابط در عالم خارج نیستند.

در بین مدل‌های دگرنمایی گراف دانش دو گروه «تطابق معنایی» و «یادگیری عمیق»، وجود نداشتن یک رابطه را معادل نبودن آن رابطه بین دو موجودیت ابتدایی و انتهایی در نظر می‌گیرند. بنابراین، این روش‌ها در مورد گراف‌های خلوت به اشتباه می‌افتند.

اما در گروه فاصله‌ی انتقالی از مدل‌های دگرنمایی گراف دانش، عملیات ریاضی انجام‌شده عملا یک کاهش بُعد ساده است و نه‌تنها در خلال دگرنمایی برداری نقطه‌ضعف ذکرشده در مورد دو گروه دیگر وجود ندارد، بلکه خروجی کاهش‌بُعدیافته در گراف‌های خلوت، یک ماتریس غنی‌تر و کاربردی‌تر نیز خواهد بود.

با توجه به مقدمه‌ی پیش‌گفته، می‌توان انتظار داشت از بین مدل‌های موجود، مدل‌های مربوط به گروه فاصله انتقالی بهترین بازده واز بین آن‌ها ساده‌ترین آن‌ها نتایج بهتری را در آزمایش‌ها رقم خواهند زد. از آنجا که سرراست[۱۸]ترین این مدل‌ها، مدل TransE می‌باشد، انتظار می‌رود که این مدل نیز بهترین نتایج را ارائه دهد.

۴- آزمایش‌ها، نتایج، و تجزیه و تحلیل

مقاله‌هایی که یک داده محک جدید را معرفی می‌کنند اصولا درصدد معرفی یک زیرمسأله جدید از یک مسأله هستند. وقتی می‌توان یک مسأله را زیرمسأله‌ی دیگری دانست که انتظار برود روش‌های حل آن دو مسأله متفاوت از هم باشند یا رفتار متفاوتی انجام دهند.

با توجه به توپولوژی متفاوت گراف‌های دانش مربوط به زبان‌های کم‌منبع همان‌طور که پیش از این اشاره شد، انتظار می‌رود روش‌های دگرنمایی گراف دانش که به منظور پیش‌بینی پیوند استفاده می‌شوند برای زبان‌های کم‌منبع رفتار متفاوتی داشته باشند.

در این بخش، مدل‌های شناخته‌شده‌ی KGE بر روی فارس‌پیش‌بین آزمایش شده تا شهودی در مورد رفتار مدل‌های KGE در دادگان محک ذکر شده ارائه شود.

همچنین بین رفتار مدل‌های فوق در این داده محک با دو دادگان محک استاندارد دیگر مقایسه می‌شوند تا تفاوت رفتارها شده، سپس تحلیل گردند.

۴-۱- جزئیات فنی آزمایش‌ها

در این بخش، مشخصات آزمایش شامل کدهای منبع استفاده شده، مدل‌های آزمایش‌شده، داده‌های ارزیابی، آزمون، و آموزش یک‌به‌یک به ترتیب توضیح داده می‌شود.

۴-۱-۱- مدل‌های آزمایش شده

در میان مدل‌های آزمایش‌شده‌ی مربوط به گروه فاصله انتقالی TransE [۳] و TransH [۱۹]، از بین مدل‌های مربوط به گروه تطبیق معنایی DistMult [۲۸] و ComplEx [۲۳]، و در نهایت از بین مدل‌های مربوط به یادگیری عمیق، مدل ConvE [۲۶] برای آزمایش استفاده می‌شوند.

۴-۱-۲- کدهای منبع[۱۹]

برای اجرا و ارزیابی مدل‌های KGE، چارچوب OpenKE[۲۰] به عنوان کتابخانه پیش‌فرض استفاده می‌شود، زیرا از اکثر مدل‌های معروف KGE پشتیبانی می‌کند. برای پیاده سازی مدل‌های ناموجود در چارچوب OpenKE، کد استفاده شده در این مقاله همان نسخه انتشاریافته توسط نویسندگان است [۲۹].

۴-۱-۳- دادگان محک مورد ارزیابی

اگرچه دادگان محکی برای زیرمسأله‌های خاص‌منظوره‌ی پیش‌بینی پیوند ارائه شده‌اند (مثل داده محک OLPBENCH [۳۰] که برای زیرمسأله‌ی «پیش‌بینی پیوند باز[۲۱]» معرفی شده است)، اما اصولاً دادگان محک که مرجع مقایسه‌ی اکثریت پژوهش‌های انجام‌شده در هر مسأله هستند، برای آن مسأله محدود می‌باشند و مسأله‌ی پیش‌بینی پیوند نیز از این مقوله مستثنی نیست.

تنها دادگان محک برای پیش‌بینی پیوند از طریق دگرنمایی گراف دانش داده محک FreeBase 15K [۷] و داده محک WordNet18 [۳۱] می‌باشند. لذا، به منظور مقایسه‌ی رفتارهای روش‌ها روی دادگان محک ارائه‌شده با دادگان محک دیگر، این دو دادگان محک استفاده می‌گردد.

۴-۱-۴- داده‌های آموزش، آزمایش و اعتبارسنجی

مجموعه‌های آموزش، اعتبارسنجی، و آزمون فارس‌پیش‌بین با روش زیر ساخته می‌شوند.

ابتدا، کل فارس‌پیش‌بین به عنوان مجموعه آموزش در نظر گرفته می‌شود. سپس، برای ساخت مجموعه‌های اعتبارسنجی و آزمون، یک مجموعه‌ی کمکی در نظر گرفته می‌شود که به تدریج از طریق رویه زیر تکمیل می‌گردد.

در این گام، به‌طور تصادفی و متناوب یکی یکی سه‌تایی‌های مجموعه آموزش انتخاب می‌شوند و هر بار اگر سه‌تایی دیگری با همین رابطه یا سه‌تایی دیگری با همین موجودیت ابتدایی در مجموعه آموزش وجود داشته باشد، آن سه‌تایی از مجموعه آموزش به مجموعه کمکی منتقل می‌شود.

این روند متناوب تا زمانی ادامه می‌یابد که اندازه‌ی مجموعه آموزش به ٪۷۰ کل فارس‌پیش‌بین کاهش یابد و به همین ترتیب اندازه مجموعه کمکی به‌اندازه ٪۳۰ کل فارس‌پیش‌بین برسد.

در نهایت، دو سوم از مجموعه کمکی به طور تصادفی به‌عنوان مجموعه‌ی آزمون و یک سوم مابقی آن به‌عنوان مجموعه اعتبارسنجی در نظر گرفته می‌شود.

۴-۲- معیارهای ارزیابی

در این بخش، معیارهای ارزیابی استفاده شده در آزمون‌های پیش‌بینی پیوند شرح داده می‌شوند. در این راستا، معیار Hit@10 ارائه شده، سپس تفاوت بین نتایج خام و فیلتر شده تشریح شده و در نهایت ارزیابی نمونه‌گیری منفی منفرد و دوگانه تفکیک شده است.

۴-۲-۱- نمونه‌گیری منفی

الگوریتم‌های پیش‌بینی پیوند به سه شکل «پیش‌بینی موجودیت‌های ابتدایی محذوف»، «پیش‌بینی روابط دانشی» و «پیش‌بینی موجودیت‌های انتهایی محذوف» قابل ارزیابی هستند؛ اما در پژوهش‌های معتبر منتشرشده اکثر ارزیابی‌ها از طریق «پیش‌بینی موجودیت‌های انتهایی محذوف» صورت می‌گیرند [۳۲].

برای پیش‌بینی یک موجودیت انتهایی محذوف، الگوریتم‌های پیش‌بینی پیوند همه موجودیت‌های گراف دانش را یک‌به‌یک به عنوان نامزد موجودیت انتهایی واقعی مورد بررسی قرار می‌دهند.

بررسی مذکور بدین‌صورت انجام می‌پذیرد که مدل KGE یک امتیاز به سه‌تایی حاصل از جایگزینی موجودیت انتهایی نامزد تخصیص می‌دهد که عملاً آن امتیاز نشان می‌دهد آن جای‌گذاری چقدر جای‌گذاری مناسبی بوده است.

به فرآیند تولید تمام سه‌تایی‌های ممکن توسط موجودیت‌های انتهایی نامزد و سپس تخصیص‌دهی امتیاز به آن‌ها و مرتب‌سازی آن‌ها برای یافتن بهترین جایگزینی فرآیند، نمونه‌گیری منفی گفته می‌شود.

۴-۲-۲- معیار Hit@10

از بین معیارهای ارزیابی موجود، معیار Hit@10 را در این مقاله استفاده می‌نماییم که نسبت دفعاتی که موجودیت صحیح در ده پیش بینی برتر ظاهر می شود را اندازه گیری می کند. در ادامه رابطه‌ی نشان‌دهنده‌ی Hit@10 را ذکر می‌نماییم.

$$Hit@10 = \frac{1}{N}\sum_{i=1}^{N}|\{p_i|rank_{correct}(p_i) \leq 10\}| \qquad (۳)$$

جایی‌که N تعداد کل پیش‌بینی‌ها، p_i لیست پیش‌بینی‌های ارائه‌شده توسط الگوریتم دگرنمایی گراف دانش، و $rank_{correct}(p_i)$ رتبه‌ی پیش‌بینی درست بین پیش‌بینی‌های ارائه‌شده در لیست p_i می‌باشد.

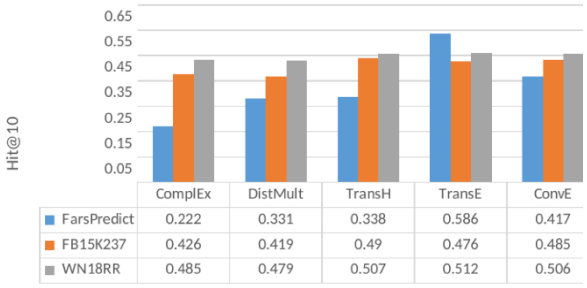
به عبارت ساده‌تر، برای هر نمونه آزمایشی، اگر موجودیت صحیح در ده پیش‌بینی برتر رتبه‌بندی شود، یک شمارش موفق به حساب می‌آید و امتیاز Hit@10 نسبت شمارش‌های موفق در کل نمونه‌های آزمایشی را بازنمایی می‌کند.

۴-۲-۳- ارزیابی خام در مقابل فیلتر شده

در رویکرد نمونه‌گیری منفی که پیش‌تر به آن اشاره شد، کل سه‌تایی‌های نامزد گراف دانش بر اساس امتیازات اختصاص داده شده مرتب می‌شوند. با این حال، رویکرد نمونه‌گیری منفی دیگری نیز وجود دارد که موجودیت‌هایی را که در نظر گرفتن آنها به عنوان دنباله، در یک سه‌تایی مثبت (سه‌تایی‌ای که قبلاً در آزمون آموزشی وجود داشته است) را حذف می‌کند.

بدیهی است که رویکرد دوم (فیلتر شده) با دقت بهتری نسبت به رویکرد اول (خام) عمل می‌کند، زیرا مشخص است که هیچ یک از سه‌تایی‌های مثبت پاسخ صحیحی نیست. اما در عین حال هر دو معیار ارزیابی خام و فیلترشده نقاط قوت و ضعف روش‌ها را از منظرهای متفاوتی نشان می‌دهند.

به طور خاص، رویکرد خام به کاربردهای واقعی-صنعتی نزدیک‌تر است چرا که در آن کاربردهای واقعی، سه‌تایی‌های مثبت از پیش مشخص نیستند. با همین استدلال، در این مقاله، نتایج خام



شکل ۲: معیار ارزیابی Hit@10، مربوط به آزمایش پنج مدل KGE شناخته شده بر روی فارس‌پیش‌بین، فری‌بیس و وردنت. نتایج نشان می‌دهند از آنجا که داده محک فارس‌پیش‌بین مربوط به زبان‌های کم‌منبع می‌باشد، مدلهای ساده‌تر Hit@10 بالاتری را حاصل می‌نمایند.

مدل‌های مختلف KGE در بین دادگان محک مختلف مقایسه می‌شوند.

۴-۲-۴- نمونه‌گیری منفرد در مقابل دوگانه منفی

اگرچه ارزیابی الگوریتم‌های پیش‌بینی پیوند بیش‌تر از طریق پیش‌بینی موجودیت انتهایی انجام می‌شود، اما می‌توان آن را از طریق پیش‌بینی موجودیت ابتدایی و پیش‌بینی رابطه نیز انجام داد.

برای ارزیابی منصفانه‌تر، در برخی موارد پیش‌بینی موجودیت ابتدایی و انتهایی توأمان انجام می‌شود و سپس ارزیابی روی مجموع آنها انجام می‌شود [۳۳].

تمرکز فقط بر روی پیش‌بینی موجودیت انتهایی، «نمونه‌گیری منفی منفرد[۲۲]» نامیده می‌شود، در حالی که در نظر گرفتن هر دوی آنها، «نمونه‌گیری منفی دوگانه[۲۳]» نامیده می‌شود.

با توجه به اینکه در فارس‌پیش‌بین تنوع موجودیت‌های ابتدایی دو برابر بیشتر از موجودیت‌های انتهایی است و بنابراین پیش‌بینی موجودیت‌های انتهایی کار ساده‌تری نسبت به پیش‌بینی ابتدایی است، می‌توان گفت که اتخاذ رویکرد نمونه‌گیری منفی دوگانه ارزیابی را منصفانه‌تر و سخت‌تر می‌کند. لذا، در این مقاله، رویکرد نمونه‌گیری منفیِ دوگانه اتخاذ شده است.

۴-۳- نتایج و بحث

در این بخش، گزارش‌های تجربی آزمایش‌های پیش‌گفته ارائه شده است. این آزمایش مدل‌های KGE ذکر شده را بین دادگان محک فارس‌پیش‌بین، FreeBase 15K-237 و WordNet 18RR مقایسه می‌کند. به عبارت دیگر، این آزمایش تأثیر پراکندگی داده‌ی محک فارس‌پیش‌بین را بر میزان صحت نتایج، در مقایسه با دو داده‌ی محک استاندارد ذکر شده نشان می‌دهد.

همانطور که در شکل ۲ مشاهده می‌شود، در بین نتایج فارس‌پیش‌بین، بهترین Hit@10 مربوط به روش‌های فاصله انتقالی است. این واقعیت با فرضیه ارائه شده در بخش ۳ مطابقت دارد. باز هم همانطور که انتظار می‌رفت، TransE با توجه به سادگی مدل و همچنین پراکندگی گراف دقیق ترین مدل است. یادآوری می‌کنیم شرح علت این مسأله در بخش ۳-۳ ذکر شده است.

همچنین در شکل ۲ تفاوت معنی‌داری بین تاب‌آوری[۲۴] مدل‌ها در فارس‌پیش‌بین در مقایسه با تلورانس آنها در فری‌بیس و وردنت است که بخاطر خلوت بودن گراف فارس‌پیش‌بین از پیش مورد انتظار بود.

۵- نتیجه‌گیری

پیش‌بینی پیوند از طریق دگرنمایی گراف دانش یک تکنیک ارزشمند برای استخراج و استنتاج دانش جدید برای تکمیل گراف دانش است.

در این پژوهش داده محکی برای پیش‌بینی پیوند مربوط به گراف‌های دانش زبان‌های کم‌منبع پیشنهاد می‌گردد. بدین‌منظور گراف دانش فارسی به‌عنوان نماینده‌ای از گراف‌های دانش کم‌منبع در نظر گرفته شده، از آنجا که خود گراف دانش فارسی (به شکل موجود)، نمی‌توانست نیازهای پیش‌بینی پیوند را برآورده کند، یک عملیات چهارمرحله‌ای برای تبدیل آن گراف به یک داده محک مناسب برای پیش‌بینی پیوند کم‌منبع ارائه گردید که برای سایر گراف‌های دانش استخراج‌شده از زبان‌های کم‌منبع نیز قابل اعمال است.

همچنین برای تحلیل نوع رفتار مدل‌های KGE روی چنین دادگان محک فرضیه‌ای با این مضمون ارائه و بحث شد که: مدل‌های ساده‌تر برای پیش‌بینی پیوند در گراف‌های دانش استخراج‌شده از زبان‌های کم‌منبع معیار صحت بالاتری را ارائه می‌دهند.

نتایج آزمایش پنج مدل استاندارد KGE روی داده محک ارائه شده، فرضیه مذکور را تایید نمود و نشان داد که مدل‌های متعلق به گروه «فاصله‌ی انتقالی» بهترین معیارهای صحت را ارائه داده، و از بین آنها نیز ساده‌ترین مدل‌ها (TransE) بهترین نرخ صحت را ارائه می‌دهد.

استفاده از داده محک ارائه شده برای محققانی که روی مساله پیش‌بینی پیوند در گراف‌های دانش پژوهش می‌کنند و به زبان‌های کم‌منبع نیز علاقه‌مند هستند، به جهت ایجاد یک بستر رقابتی برای ارائه‌ی مدل‌های با نرخ Hit@10 بالاتر توصیه می‌شود.

به‌عنوان پژوهش‌های آتی، پیشنهاد می‌کنیم سایر معیارهای ارزیابی

پیش‌بینی پیوند مانند میانگین رتبه متقابل (MRR[25]) روی داده محک مذکور آزموده و تحلیل شوند.

ملحق کردن داده محک ارائه شده به زیرساخت چندزبانه Schema.org نیز به عنوان یک طبقه بندی مرجع، یکی دیگر از کارهای آتی پیشنهادی است.

مراجع

[1] H. Yu, H. Li, D. Mao, and Q. Cai, "A domain knowledge graph construction method based on Wikipedia," *J. Inf. Sci.*, vol. 47, no. 6, pp. 783–793, 2021

[2] M. Yahya, J. G. Breslin, and M. I. Ali, "Semantic web and knowledge graphs for industry 4.0," *Appl. Sci.*, vol. 11, no. 11, pp. 5110, 2021.

[3] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Advances in Neural Information Processing Systems*, Neural information processing systems foundation, 2013.

[4] M. Shamsfard and Y. Ghazanfari, "Augmenting farsnet with new relations and structures for verbs," in *Proceedings of the 8th Global WordNet Conference, GWC 2016*, 2016, pp. 377–383.

[5] M. Asgari-Bidhendi, A. Hadian, and B. Minaei-Bidgoli, "FarsBase: The Persian knowledge graph," *Semant. Web*, vol. 10, no. 6, pp. 1169–1196, 2019.

[6] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A Nucleus for a Web of Open Data," in *The Semantic Web*, K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, and P. Cudré-Mauroux, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 722–735, 2007.

[7] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *In SIGMOD Conference*, pp. 1247–1250, 2008.

[8] R. Navigli and S. P. Ponzetto, "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *Artif. Intell.*, vol. 193, pp. 217–250, 2012.

[9] D. Vrandečić and M. Krötzsch, "Wikidata: A free collaborative knowledgebase," *Commun. ACM*, vol. 57, no. 10, pp. 78–85, 2014.

[10] T. Pellissier Tanon, G. Weikum, and F. Suchanek, "YAGO 4: A Reason-able Knowledge Base," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* , pp. 583–596, 2020.

[11] A. Uyar and F. M. Aliyu, "Evaluating search features of Google Knowledge Graph and Bing Satori: Entity types, list searches and query interfaces," *Online Inf. Rev.*, vol. 39, no. 2, pp. 197–213, 2015.

[12] Z. Wang *et al.*, "XLore: A large-scale English-Chinese bilingual knowledge graph," in *CEUR Workshop Proceedings*, pp. 121–124, 2013.

[13] E. Montiel-Ponsoda and V. Rodr'iguez-Doncel, "Lynx: Building the legal knowledge graph for smart compliance services in multilingual Europe," in *Proceedings of the 1st workshop on LREC (language resources and technologies for the legal knowledge graph) workshop*, pp. 19–22, 2018.

[14] J.-P. Prost, "Holinet: Holistic knowledge graph for french," *Journées Jointes des Group. Rech. Linguist. Informatique, Formelle Terrain*, vol. 1, pp. 123–130, 2022.

[15] S. Y. Yoo and O. R. Jeong, "Automating the expansion of a knowledge graph," *Expert Syst. Appl.*, vol. 141, p. 112965, 2020.

[16] M. Mirzapour *et al.*, "Introducing RezoJDM16k: a French KnowledgeGraph DataSet for link prediction," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 5163–5169, 2022.

[17] H. Alberts *et al.*, "VisualSem: A High-quality Knowledge Graph for Vision & Language," in *MRL 2021 - 1st Workshop on Multilingual Representation Learning, Proceedings of the Conference*, pp. 138–152, 2021.

[18] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Proceedings of the National Conference on Artificial Intelligence*, AI Access Foundation, pp. 2181–2187, 2015.

[19] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *Proceedings of the National Conference on Artificial Intelligence*, AI Access Foundation, pp. 1112–1119, 2014.

[20] Z. Sun, Z. H. Deng, J. Y. Nie, and J. Tang, "Rotate: Knowledge graph embedding by relational rotation in complex space," in *7th International Conference on Learning Representations, ICLR 2019*, pp. 1-18, 2019.

[21] M. Nickel, V. Tresp, and H. P. Kriegel, "A three-way model for collective learning on multi-relational data," in *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pp. 809–816, 2011.

[22] M. Nickel, L. Rosasco, and T. Poggio, "Holographic embeddings of knowledge graphs," in *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, pp. 1955–1961, 2016.

[23] T. Trouillon, J. Welbl, S. Riedel, E. Ciaussier, and G. Bouchard, "Complex embeddings for simple link prediction," in *33rd International Conference on Machine Learning, ICML 2016*, pp. 3021–3032, 2016.

[24] B. Shi and T. Weninger, "ProjE: Embedding projection for knowledge graph completion," in *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, pp. 1236–1242, 2017.

[25] D. Q. Nguyen, D. Q. Nguyen, T. D. Nguyen, and D. Phung, "A convolutional neural network-based model for knowledge base completion and its application to search personalization," *Semant. Web*, vol. 10, no. 5, pp. 947–960, 2019.

[26] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel, "Convolutional 2D knowledge graph embeddings," in *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pp. 1811–1818, 2018.

[27] C. Shang, Y. Tang, J. Huang, J. Bi, X. He, and B. Zhou, "End-to-end structure-aware convolutional networks for knowledge base completion," in *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pp. 3060–3067, 2019.

[28] B. Yang, W. tau Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," in *3rd International Conference on Learning Representations, ICLR 2015 Conference Track Proceedings*, pp. 1-12, 2015.

[29] X. Han *et al.*, "OpenKE: An open toolkit for knowledge embedding," in *EMNLP 2018 - Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Proceedings*, pp. 139–144, 2018.

[30] S. Broscheit, K. Gashteovski, Y. Wang, and R. Gemulla, "Can we predict new facts with open knowledge graph embeddings? A benchmark for open link prediction," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 2296–2308, 2020.

[31] C. Fellbaum, "WordNet," in *Theory and Applications of*

[32] M. Wang, L. Qiu, and X. Wang, "A survey on knowledge graph embeddings for link prediction," *Symmetry (Basel).*, vol. 13, no. 3, 2021.

[33] P. Rosso, D. Yang, and P. Cudré-Mauroux, "Beyond Triplets: Hyper-Relational Knowledge Graph Embedding for Link Prediction," in *The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020*, pp. 1885–1896, 2020.

پاورقی‌ها:

[1] Head entity
[2] Tail entity
[3] Knowledge Graph Embedding
[4] Translational distance models
[5] Semantic matching models
[6] Deep learning models
[7] Benchmark
[8] FarsBase
[9] Automatic expansion knowledge graph
[10] Tensor
[11] Latent semantic
[12] Interaction matrix
[13] Link
[14] Ontologic
[15] Attribute
[16] Value
[17] FreeBase
[18] Straightforward
[19] Source-codes
[20] https://github.com/thunlp/OpenKE
[21] Open link prediction
[22] Single negative sampling
[23] Double negative sampling
[24] Tolerance
[25] Mean Reciprocal Rank (MRR)