Biannual scientific-research journal of language and linguistics
Fall & Winter 2025, 20 (40), 131-152
https://lsi-linguistics.ihcs.ac.ir/
DOI: 10.30465/LSI.2025.50863.1791

Original Research
Received: 01/Dianec/2025 *** Accepted: 14/Mar/2025
ISSN: 2322-3847

Knowledge Graph Embedding for Improving Persian Question Answering

Ali Asad[1] Computer Engineering Department, Amirkabir University of Technology, Tehran, Iran

Saeedeh Momtazi[2] Computer Engineering Department, Amirkabir University of Technology, Tehran, Iran

Abstract

Knowledge graphs (KGs) represent structured data through triples that encode entities and their relationships. While they are traditionally used for storing explicit facts, embedding these triples into a vector space enables the discovery of implicit relations, enhancing reasoning in tasks such as question answering (QA). This study explores the challenge of answering Persian-language questions using Fars-Wiki-KG, a large-scale knowledge graph automatically constructed from Persian Wikipedia. We create a QA system that combines the knowledge graph and natural language questions into the same vector space, allowing for complete reasoning on both simple and complex queries. The system employs ComplEx for graph embeddings and XLM-RoBERTa, a multilingual transformer model, for question embeddings. Additionally, we introduce a Persian QA dataset containing over 74,000 question-answer pairs generated from Fars-Wiki-KG. Experimental results show that our framework effectively handles complex reasoning in Persian, with strong performance on both simple and multi-hop questions.

Keywords: knowledge graph embedding, knowledge graph-based question answering, Persian knowledge graph.

1. aliasad059@aut.ac.ir
2. momtazi@aut.ac.ir (Corresponding Author)
How to cite: Asad, A., & Momtazi, S. (2024). Knowledge Graph Embedding for Improving Persian Question Answering. *Language and Linguistics*, 20(40), 131- 152. doi: 10.30465/LSI.2025.50863.1791

1. Introduction

By storing data in knowledge graphs, it is possible to retrieve not only explicit relationships but also implicit ones, as processing the entire graph at once can reveal hidden facts not explicitly present in the original graph (Choudhary et al., 2021). This capability enables question-answering systems to go beyond the information provided by knowledge graph triples and tackle more complex challenges. Knowledge graph embeddings, which represent the nodes and edges of the graph as numerical vectors, are developed to aid this purpose. These embeddings are created by considering the entire graph, so what these numbers represent is more than the explicit facts of the knowledge graph. For example, we can ascertain an individual's profession through indirect indicators such as their authored works. Embedding the graph into a vector space allows for uncovering such hidden relations by modeling semantics at a global level. The technique will also help address the issues of incompleteness and sparsity in the knowledge graphs, which are inevitable. This paper explores the task of answering Persian questions using embeddings derived from a Persian knowledge graph called Fars-Wiki-KG (Shirmardi et al., 2021).

2. Literature Review

2.1. Related Work on Embedding Models

Knowledge graph embedding models divide into two main approaches: translation-based models and semantic models. Translation-based models like TransE show relationships as movements in a vector space (Bordes et al., 2013), while RotatE builds on this by treating relationships as rotations in a complex space, which helps better manage symmetric and inverse relationships (Sun et al., 2019). Semantic models such as DistMult (Yang et al., 2015) and ComplEx (Trouillon et al., 2016) score triples based on similarity. ComplEx, by operating in complex spaces, effectively models asymmetric relationships and is widely adopted in QA tasks.

2.2. Related Work on QA over Knowledge Graphs

Recent QA systems embed both the question and KG entities into a shared space. The question embedding acts as a latent relation are used to retrieve answer entities. This approach, introduced by Saxena et al. (2020), allows multi-hop reasoning without explicit path traversal.

3. Method

Fars-Wiki-KG (Shirmardi et al., 2021) contains well over two million nodes and 7.5 million triples, covering over 90% of the Persian Wikipedia infoboxes. Based on the nodes and edges of this knowledge graph, 1370 patterns have been created to build a question-answering dataset. These

patterns include both simple and multi-hop questions. For example, a simple pattern might be "Which country is the producer of X?" In contrast, a multi-hop pattern could be: "What language do the people speak in the country of the producer of X?" The first one is straightforward, while the second question involves two hops: first finding the country of X, and then determining the language associated with that country. By considering these patterns and extracting their corresponding answers from Fars-Wiki-KG, we provided a dataset consisting of questions and answers. Having a question like "Which country is the producer of 'Breaking Bad'?", the head entity is 'Breaking Bad,' the relation is the country of origin, and the tail entity is the 'USA,' which is the answer.

Various knowledge graph embedding models, such as translational and semantic matching models, are trained and evaluated on subsets of the knowledge graph to embed its nodes. This process captures both explicit and latent information between the nodes in numerical vector form. At the same time, Persian questions are embedded using a multilingual version of the BERT language model, namely XLM-RoBERTa (Liu et al., 2019), to represent the relationships between each question and its corresponding answer. Each question starts with a head entity, followed by a relationship that describes how to reach the tail, which is the answer**.** By utilizing the head entity embedding and the relation embedding (which contains information about the question), a score is assigned to all other nodes in the graph. This score indicates the likelihood of each node being the tail entity (question's possible answers). Therefore, the above-explained approach, inspired **by** the research of Saxena et al. (2020), allows the system to answer questions even when the relevant triples are not explicitly included in the graph. Moreover, it enables responses to more complex questions that require traversing multiple edges within the graph, called multi-hop questions.

4. Results

To evaluate our model, we provided a dataset of simple and complex question-answer pairs for Persian, containing over 74,000 question-answer pairs derived from this graph. Table 1 presents the results of our proposed method.

Table 1

Accuracy and Hit@10 Results on Persian QA Dataset

Dataset	Size	Accuracy (%)	H@10 (%)
Simple Questions	60,630	92	96
Complex Questions	13,787	54	82
All (Combined)	74,417	85	94

The suggested framework shows around 85% accuracy and an H@10 of 94% on a dataset made up of both simple and multi-step Persian question-answer pairs, which suggests good results for the question-answering task in the Persian knowledge graph.

6. Conclusion

We showed a Persian QA system that uses KG embedding, proving that even with few available KGs, embedding methods can produce good results. By matching Fars-Wiki-KG entity embeddings with question vectors from XLM-RoBERTa, our system learns to find answers without needing exact matches. This is especially useful for multi-hop questions where we need to figure out intermediate entities without them being directly stated. By matching the Fars-Wiki-KG entity representations with question vectors from XLM-RoBERTa, our system can figure out answers without needing exact matches of data triples. This technique is particularly valuable for multi-hop questions where intermediate entities must be inferred implicitly. Future directions include extending deeper graph reasoning, optimizing model architectures, and evaluating generalization across more diverse and ambiguous questions.

Acknowledgments

This work is based upon research funded by Iran National Science Foundation (INSF) under project No. 4031813.

مقاله پژوهشی

تاریخ دریافت: ۱۴۰۳/۱۰/۱۲ *** تاریخ پذیرش: ۱۴۰۳/۱۲/۲۴

ISSN: ۲۳۲۲-۳۸۴۷

دوفصلنامه علمی- پژوهشی زبان و زبان‌شناسی
دوره ۲۰، شماره ۴۰، پاییز و زمستان ۱۴۰۳، ۱۳۱- ۱۵۲
lsi-linguistics.ihcs.ac.ir
DOI: 10.30465/LSI.2025.50863.1791

تعبیه گراف دانش به منظور بهبود سامانه‌های پرسش و پاسخ فارسی

علی اسد | کارشناسی مهندسی کامپیوتر، دانشگاه امیرکبیر، تهران، ایران

سعیده ممتازی | دانشیار دانشکده مهندسی کامپیوتر دانشگاه امیرکبیر، تهران، ایران

چکیده

با ذخیره داده‌ها در گراف‌های دانش می‌توان علاوه بر روابط صریح، روابطی ضمنی را نیز حدس و بازیابی کرد. این ویژگی به سامانه‌های پرسش و پاسخ این امکان را می‌دهد که فراتر از آنچه از سه‌گانه‌های گراف دیده‌اند به چالش کشیده شوند. تعبیه گراف دانش یا نمایش گره‌ها و یال‌های گراف در قالب بردارهای عددی، به همین منظور صورت می‌پذیرد. در این مقاله مسئله پاسخ به پرسش‌های فارسی با استفاده از تعبیه گراف دانش فارسی بررسی شده است. مدل‌های مختلفی برای تعبیه گراف دانش آموزش داده شده‌اند تا هویت پیدا و نهان گره‌ها را در قالب بردار بازنمایی کند. از سوی دیگر با استفاده از مدل های زبانی، پرسش‌های فارسی به گونه‌ای تعبیه می‌شوند که نمایانگر یال ضمنی و یا عینی بین هر پرسش و پاسخ مربوطه باشد. با این رویکرد می‌توان به پرسش‌هایی پاسخ داد که مستقیماً سه‌گانه مربوطه در گراف آورده نشده است و همچنین پا را فراتر گذاشته و به پرسش‌های پیچیده‌تر که نیازمند طی چندین یال است نیز، پاسخ مناسب داد. نتایج حاصل از مدل پیشنهادی مبتنی بر تعبیه گراف دانش فارسی فارس-ویکی-کی‌جی برای پاسخ‌گویی به پرسش‌های فارسی، نشان‌دهنده دقت ۸۵ درصد بر روی مجموعه‌داده پرسش و پاسخ ساده و پیچیده به زبان فارسی می‌باشد.

کلیدواژه: گراف دانش فارسی، تعبیه گراف دانش، سامانه پرسش و پاسخ مبتنی بر گراف دانش.

۱. مقدمه

گراف دانش[۱] نمایشی ساختاریافته از داده‌ها است که می‌تواند حقایق را در قالب گره‌ها[۲] و یال‌های[۳] مختلف ذخیره و بازیابی کند (Choudhary et al., 2021). در این گراف، هر گره نمایانگر یک موجودیت[۴] است و هر یال مشخص می‌کند این موجودیت‌ها با یکدیگر چه رابطه‌ای[۵] برقرار می‌کنند. هر موجودیت می‌تواند با چندین یال به موجودیت‌های دیگر متصل شود و بدین صورت، تمام اطلاعاتی که از هر موجودیت داریم را می‌توانیم در قالب گراف، به‌صورت ساختارمند ذخیره کنیم. هر حقیقت[۶] در گراف دانش، در قالب سه‌گانه‌هایی ذخیره می‌شوند که نمایانگر دو موجودیت ابتدایی و انتهایی و نیز رابطه بین آنها است. مثلا سه‌گانه (سعدی، آرامگاه، سعدیه) مشخص می‌کند بین «سعدی» به عنوان موجودیت ابتدایی و «سعدیه» به عنوان موجودیت انتهایی، رابطه «آرامگاه» برقرار است. گراف‌های دانش مختلفی در زبان انگلیسی ساخته و پرداخته شده‌اند. از جمله آنها می‌توان به گراف‌های دانش فری‌بیس[۷] (Bollacker et al., 2008)، دی‌بی‌پدیا[۸] (Lehmann et al., 2015) وهمچنین ویکی‌دیتا[۹] (Vrandečić et al., 2014) اشاره داشت. در زبان فارسی نیز گراف‌های دانش فارس‌بیس[۱۰] (Asgari-Bidhendi et al., 2019) و فارس-ویکی-کی‌جی[۱۱] (Shirmardi et al., 2021) موجود هستند.

به دلیل توانمندی ذخیره داده‌ها به‌صورت ساختارمند و قابل تفسیر برای ماشین‌ها، می‌توان در کاربردهای متعددی از گراف‌های دانش استفاده کرد. از جمله آنها می‌توان به پیش‌بینی پیوند، دسته‌بندی سه‌گانه‌ها، دسته‌بندی موجودیت‌ها و یا حتی کاربردهای متداول‌تر در استخراج رابطه، سامانه‌های پرسش و پاسخ و سامانه‌های توصیه‌گر اشاره داشت (Rossi et al., 2021; Wang et al., 2017).

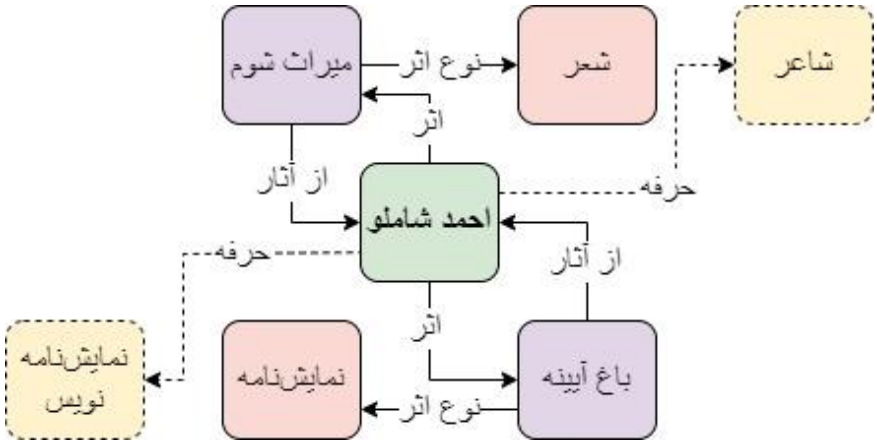
با وجود تمامی پژوهش‌های موفقی که در پردازش زبان طبیعی و درک زبان انجام شده است، همچنان چالش‌هایی در زمینه استنباط اطلاعاتی که از پیش به‌صورت مستقیم بیان نشده است، وجود دارد. مثلاً در جمله «احمد شاملو خالق دو اثر میراث شوم و باغ آیینه است»،

---

1. knowledge graph
2. node
3. edge
4. entity
5. relationship
6. fact
7. FreeBase
8. DBpedia
9. WikiData
10. FarsBase
11. FarsWikiKG

با استفاده از مدل‌های زبانی به سادگی نمی‌توان فهمید میراث شوم نام یک نمایشنامه و دیگری نام یک مجموعه شعر است و این‌گونه استدلال کرد که احمد شاملو نویسنده، فیلم‌نامه‌نویس و شاعر ایرانی بوده است به عبارتی حرفه شخص را استنباط کرد. اما به وسیله ساختار گراف دانش، در ارتباط با موجودیت «احمد شاملو»، بازیابی حرفه وی تنها با گردش در یال‌ها و موجودیت‌های همسایه، به سادگی امکان‌پذیر است. آنچه زاید مشکل در درک متن و به‌صورت مشابه، پاسخ به پرسش‌ها با استفاده از گراف دانش است، نبودن یال «حرفه» به‌صورت مستقیم در کنار موجودیت «احمد شاملو» است. همانطور که پیشتر گفته شد، مشکل مطرح شده که از ناقص بودن گراف دانش بر می‌آید، تا حد خوبی با تعبیه گراف دانش قابل حل است (Saxena et al., 2020). حتی اگر یال «حرفه» به‌صورت مستقیم در کنار این موجودیت نباشد، از ارتباط این موجودیت با آثار خود، و ویژگی‌ها و یال‌های مجاور هر کدام از این آثار، می‌توان حرفه وی را به‌صورت ضمنی برداشت کرد. شکل ۱ مقصود از آنچه گفته شد را به تصویر می‌کشد.

شکل ۱- چالش‌های پاسخ به پرسش‌های چندپرشی در گراف دانش



در این مقاله با استفاده از تعبیه گراف دانش فارسی به پاسخ‌گویی سوالات در یک سامانه پرسش و پاسخ فارسی پرداخته‌ایم. در این رویکرد با تعبیه پرسش و گراف سعی می‌نماییم پاسخ سوالات ساده و پیچیده را به یک معماری انتها-به-انتها بسپاریم. در بخش دوم به مروری بر کارهای پیشین درباره مدل‌های تعبیه گراف دانش و سامانه‌های پرسش و پاسخ می‌پردازیم و رویکردهای پیشنهادی مختلفی را مورد بررسی قرار می‌دهیم. بخش سوم به معرفی گراف دانش فارس-ویکی-کی‌جی و نیز مجموعه داده پرسش و پاسخ فارسی برآمده از این گراف دانش می‌پردازد. در بخش چهارم معماری پیشنهادی این پژوهش توضیح داده خواهد شد. در بخش پنجم آزمایش‌ها، نتایج و ارزیابی‌های مختلف انجام شده بر روی کیفیت مدل‌های تعبیه

گراف دانش و همچنین عملکرد سامانه پرسش و پاسخ فارسی با استفاده از گراف دانش مورد بحث قرار می‌گیرند. در بخش ششم جمع‌بندی و نیز پیشنهادهایی برای کارهای پیشرو، به منظور بهبود عملکرد سامانه پرسش و پاسخ، بیان می‌گردد.

۲. مروری بر کارهای پیشین

در کاربردهای بیان‌شده، تعامل و بازیابی سه‌گانه‌های درون گراف دانش می‌تواند به وسیله درخواست‌های منطقی صورت پذیرد؛ اما نمی‌توان تنها به این داده‌ها اکتفا کرد. علی‌رغم داده‌های فراوانی که به‌صورت مستقیم در گراف دانش وجود دارند، این نوع ساختارها همچنان از ناقص بودن[1] و تنک بودن[2] رنج می‌برند (Rossi et al., 2021; Wang et al., 2017; Bordes et al., 2013) به عبارتی تعداد زیادی از یال‌های فراموش شده در ساختار گراف موجود هستند که به‌صورت مستقیم بیان نشده‌اند. اینجاست که تعبیه گراف دانش در قالب نمایش‌های عددی می‌تواند بسیار مفید باشد. ایده و انگیزه اصلی تعبیه گراف دانش این است که اجزای تشکیل‌دهنده گراف دانش را به یک فضای برداری برده و به گونه‌ای نمایش دهیم که انجام عملیات‌های مختلف را ساده‌تر کند (Rossi et al., 2021). به بیانی دیگر به ازای هر موجودیت و هر یال، یک بازنمایی عددی (معمولاً با ابعاد کوچک) ساخته شود به گونه‌ای که دانش و حقایق ذخیره‌شده در گراف دانش اولیه همچنان محفوظ بماند. مزیت اصلی این نوع نمایش علاوه بر کوچک بودن ابعاد این بردارهای عددی، این است که طی فرآیند ساخت و یادگیری این بردارها، به ازای هر یک از گره‌ها و یال‌های منفرد، کل اجزای گراف در تعامل با هم قرار می‌گیرند و حقایقی که به‌صورت مستقیم در ساختار گراف دانش اولیه قرار ندارند نیز کشف می‌شود (Huang et al., 2019). در این قسمت ابتدا به بررسی مدل‌های تعبیه گراف دانش می‌رویم و سپس سامانه‌های پرسش و پاسخ مبتنی بر آن را مرور می‌کنیم.

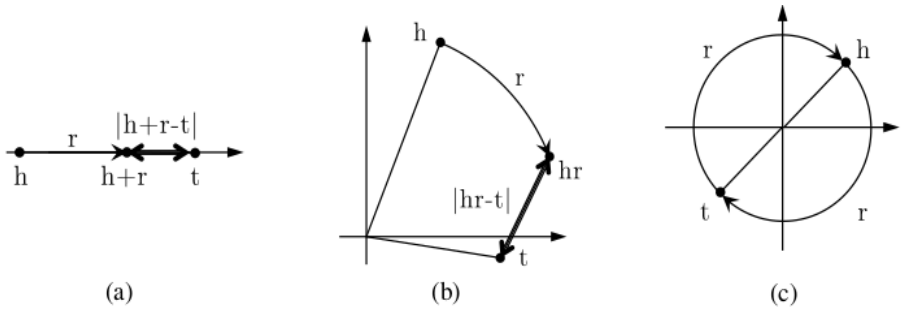
۲-۱. مروری بر مدل‌های تعبیه گراف دانش

به‌طور کلی این مدل‌ها به دو دسته کلی مدل‌های ترجمه‌ای[3] و مدل‌های تطبیق معنایی[4] تقسیم می‌شوند. در مدل‌های ترجمه‌ای از سنجه‌های فاصله‌ای برای به‌دست آوردن امتیاز شباهت یک جفت موجودیت و روابط بین آنها استفاده می‌شود. این مدل‌ها تلاش می‌کنند تا موجودیت‌ها را به فضاهای مختلف برده و با استفاده از تابع امتیازدهی مختص خود، برای موجودیت‌های

1. incompleteness
2. sparsity
3. translational model
4. semantic matching model

تبدیل شده، نمایشی از بردارهای عددی تولید کنند (Choudhary et al., 2021). دو مدل ترنس‌ای[1] و روتیت‌ای[2] از دسته مدل‌های ترجمه‌ای هستند. اولین مدلی که بر این پایه معرفی شد، مدل ترنس‌ای توسط بوردس[3] در سال ۲۰۱۳ بود (Bordes et al., 2013). این مدل به ازای هر جزء سه‌گانه در گراف دانش، یک بردار می‌سازد به گونه‌ای که حاصل جمع بردار موجودیت ابتدایی و بردار رابطه به موجودیت انتهایی برسد. در سال ۲۰۱۹ مدل روتیت‌ای توسط سان[4] معرفی شد (Sun et al., 2019). در این مدل برخلاف ترنس‌ای، از فضای قطبی برای نمایش و تبدیل بردارها استفاده می‌شود. به‌صورت مشخص‌تر سان با ایده گرفتن از رابطه اویلر، بیان کرد که در روتیت‌ای، بهتر است تبدیل موجودیت ابتدایی به انتهایی به‌صورت یک چرخش در صفحه فضای مختلط مدل شود تا به‌صورت یک تبدیل ساده خطی در فضای اقلیدسی (Sun et al., 2019). در شکل ۲ مقایسه بین مدل ترنس‌ای و روتیت‌ای به‌صورت بصری آورده شده است.

شکل ۲- نمایش عملکرد ترنس‌ای و روتیت‌ای با تعبیه‌های یک بعدی



(a) مدل ترنس‌ای تبدیل r را بر روی یک خط انجام می‌دهد. (b) مدل روتیت‌ای تبدیل r را با چرخش در یک فضای مختلط انجام می‌دهد. (c) روتیت‌ای: مثالی برای شبیه‌سازی رابطه متقارن.

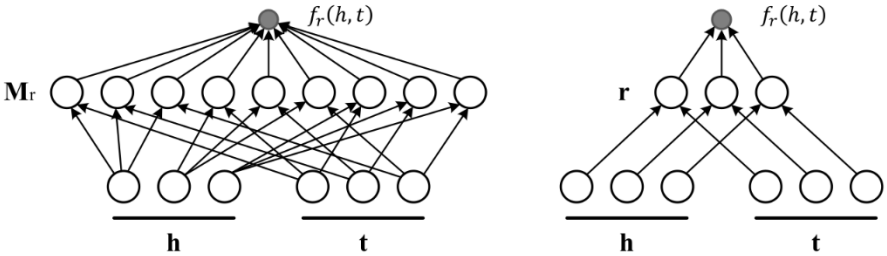
مدل‌های مبتنی بر تطبیق معنایی از توابع امتیازدهی مبتنی بر شباهت استفاده می‌کند. به بیان دیگر درستی حقایق مختلف با تطبیق معنایی نهفته موجودیت‌ها و روابطی که در فضای برداری تعبیه شده‌اند اندازه‌گیری می‌شود (Wang et al., 2017). مدل‌های رسکل[5]، دیست‌مالت[6] و کامپلکس[7] از این دسته هستند.

1. TransE
2. RotatE
3. A. Bordes
4. Z. Sun
5. RESCAL
6. DistMult
7. ComplEx

مدل رسکل مدلی بر پایه یادگیری رابطه‌ای آماری[1] است و با استفاده از تجزیه ماتریسی[2] عمل می‌کند. در این مدل سه‌گانه‌های موجود در گراف دانش در یک تنسور سه‌مسیره قرار می‌گیرند. دو بعد از این تنسور موجودیت‌های ابتدایی و انتهایی، و بعد سوم نیز رابطه‌ها را نگه می‌دارد (Nickel et al., 2011). مدل دیست‌مالت مدل ساده‌شده رسکل است که از تابع امتیازدهی متفاوتی استفاده می‌کند (Yang et al., 2015). با استفاده از این تابع امتیازدهی، پیچیدگی محاسباتی و فضای ذخیره‌سازی در این مدل بهتر از رسکل شد. اما با توجه به ساده‌سازی انجام شده در آن، تنها برای روابط متقارن کارکرد قابل قبولی دارد (Wang et al., 2017).

شکل ۳ نمایشی از مقایسه تابع امتیازدهی مدل‌های رسکل و دیست‌مالت است.

شکل ۳- مقایسه روش امتیازدهی رسکل و دیست‌مالت



(a) RESCAL. (b) DistMult.

وجود روابط نامتقارن در گراف‌های دانش مشکلی بود که در تعبیه‌های ساخته‌شده در مدل‌های قبلی وجود داشت. مدل کامپلکس به رفع این مشکل پرداخت. موجودیت‌ها بسته به اینکه در ابتدا یا انتهای سه‌گانه قرار می‌گیرند ممکن است مفهوم متفاوتی را برسانند. این دو موجودیت در هر حالت ابتدایی یا انتهایی نقش‌های متفاوتی دارند و برای رفع این مشکل باید به ازای هر نقش در این سه‌گانه برداری مجزا ساخته شود. اما این کار باعث افزایش تعداد پارامترهای مدل می‌شود. تعبیه‌های ساخته شده در مدل کامپلکس با یادگیری توأم موجودیت‌ها در هر دو نقش، از رابطه‌های نامتقارن نیز پشتیبانی می‌کند. در این مدل از ضرب نقطه‌ای هرمیشن بین موجودیت‌های ابتدایی و انتهایی استفاده می‌شود. علاوه بر آن، یک ماتریس قطری مرتبه پایین‌تر با تجزیه بردارهای ویژه ساخته شده و با استفاده از آن، پیش‌بینی

1. statistical relational learning
2. matrix factorization

درستی سه‌گانه‌ها انجام می‌شود (Trouillon et al., 2016). جدول ۱ مقایسه‌ای از تابع امتیازدهی مدل‌های گفته شده و نیز روابطی که پشتیبانی می‌کنند دارد.

جدول ۱- مقایسه مدل‌های تعبیه گراف دانش

مدل	تابع امتیازدهی	متقارن	نامتقارن	وارون	ترکیب	یک به چند
ترنس‌ای	$-\|\|h + r - t\|\|$	-	+	+	+	-
روتیت‌ای	$-\|\|h \circ r - t\|\|$	+	+	+	+	+
رسکل	$h^T M_r t$	+	+	+	-	+
دیست‌مالت	$h^T diag(r) t$	+	-	-	-	+
کامپلکس	$Re(h^T diag(r) t)$	+	+	+	-	+

۲.۲. مروری بر سامانه‌های پرسش و پاسخ مبتنی بر تعبیه گراف دانش

در سامانه‌های پرسش و پاسخ با استفاده گراف دانش، یک پرسش به زبان طبیعی مطرح شده و پاسخ صحیح بر اساس ادراک سامانه از پرسش و از درون گراف دانش استخراج می‌شود. یک پرسش گاهی به‌صورت‌های مختلفی بیان می‌شود و می‌تواند چندین پاسخ داشته باشد. به علاوه، برای پاسخ‌دهی به پرسش با استفاده از گراف دانش، ممکن است نیازمند پیمایش چندین یال روی گراف باشیم. همه این چالش‌ها باعث می‌شود روش‌های مختلفی برای پاسخ‌دهی به پرسش‌ها مبتنی بر گراف دانش ارائه شود. به عنوان مثال در روشی موسوم به پاسخ‌گویی جعبه‌ای[1]، فرض بر این است که پرسش‌ها در فضایی محدود به شکل جعبه تعبیه می‌شود و پاسخ‌ها به‌صورت منطقی از درون این جعبه‌ها استخراج می‌شوند. در این روش، عملیات‌های منطقی مانند "و"[2] و "یا"[3] بر روی جعبه‌ها اعمال می‌شود و پاسخ‌ها در فضای خروجی این عملیات‌های منطقی قرار می‌گیرند (Ren et al., 2020). برای تسکین مشکل ناقص بودن گراف‌های دانش، روش گرفت‌نت، با استفاده از هر دو منبع ساختارمند (گراف دانش) و غیرساختارمند (متن)، به ازای هر پرسش یک زیرگراف می‌سازد که داده‌های غنی‌تری را نسبت به گراف اولیه ذخیره می‌کند (Sun et al., 2018). در روش پولنت[4] که اقتباسی از ایده گرفت‌نت است، با استفاده از شبکه‌های عصبی پیچشی گرافی، زیرگراف‌های بهینه‌تری برای پاسخ به پرسش‌ها ایجاد می‌کند (Sun et al., 2019). در نهایت، در روش تعبیه گراف دانش و پرسش و پاسخ[5] که معماری اصلی استفاده شده در این مقاله نیز است، از مدل‌های تعبیه

1. Queery2Box
2. AND
3. OR
4. PullNet
5. EmbedKGQA

گراف دانش برای تبدیل موجودیت‌های گراف به نمایش‌های عددی استفاده می‌کند. سپس، از مدل‌های زبانی عمیق که توانایی بالایی در درک متن دارند، برای تعبیه پرسش‌ها بهره می‌برد. خروجی این مدل‌ها به فضایی مشترک منتقل می‌شود، جایی که درک پرسش به‌صورت یک مسیر فرضی از موجودیت آغازین به موجودیت انتهایی (پاسخ پرسش) نمایش داده می‌شود. در نهایت، با استفاده از توابع امتیازدهی، موجودیت‌هایی که بیشترین امتیاز را دارند به عنوان پاسخ نهایی باز می‌گردند (Saxena et al., 2020).

در بخش‌های بعدی این معماری به‌طور کامل تشریح و برای استفاده در گراف دانش فارسی تطبیق داده خواهد شد.

۳. دادگان

۳ .۱. گراف دانش فارسی

در این بخش به معرفی گراف دانش فارس-ویکی-کی‌جی که برآمده از آزمایشگاه پردازش زبان طبیعی دانشکده مهندسی کامپیوتر دانشگاه صنعتی امیرکبیر است می‌پردازیم (Shirmardi et al., 2021). همچنین به اقداماتی که جهت بهره‌وری از گراف دانش در این پژوهش استفاده شده است اشاره خواهد شد.

این گراف دانش با استفاده از داده‌ها و پیوندهای موجود در جداول اطلاعات[1] صفحات دانش‌نامه ویکی‌پدیا ساخته شده است و این قابلیت را دارد که به‌صورت خودکار بروزرسانی شود. داشتن این ویژگی کمک می‌کند علی‌رغم هر گونه تغییر در داده‌های موجود در ویکی‌پدیا که معمولا دور از انتظار نیز نیست، ساختار گراف دانش بروزرسانی شده و داده‌های آن، پایدارتر و درست‌تر باشند. این گراف دانش با پوشش ۹۰ درصدی بر روی صفحات ویکی‌پدیا، دربرگیرنده بیش از ۲ میلیون موجودیت یکتا و ۷/۵ میلیون حقیقت درباره ارتباط بین موجودیت‌ها است. فارس-ویکی-کی‌جی به قالب RDF[2] ذخیره شده است و با اجرای درخواست‌های اسپارکل[3]، می‌تواند به بازیابی داده‌های درون گراف دانش بپردازد (Shirmardi et al., 2021). در این پژوهش گراف دانش در سامانه مدیریت پایگاه‌داده نئو۴جی[4] ذخیره و استفاده شده است. در ادامه گزارش برای پرهیز از تکرار اسم، در نظر داشته باشید که منظور از گراف دانش فارسی، گراف دانش فارس-ویکی-کی‌جی است.

1. infobox
2. Resource Description Framework(RDF)
3. SPARQL
4. Neo4j

۳. ۲. مجموعه داده پرسش و پاسخ روی گراف دانش

در این پژوهش مجموعه داده پرسش و پاسخی ساخته و پرداخته شده است که جواب‌ها، مستقیماً از داخل گراف دانش فارسی بدست آورده شده باشند. این مجموعه پرسش و پاسخ، شامل پرسش‌های ساده و پرسش‌های پیچیده (چندپرشی[1]) می‌شود. در پرسش‌های ساده، ۱۳۷۰ الگو[2] برای ۶۰ رابطه یکتا در گراف دانش در نظر گرفته شده است. هر الگو رابطه را به زبان طبیعی بیان می‌کند. مثلا به ازای رابطه «کشور سازنده» ۲۰ الگوی متفاوت در نظر گرفته شده است. مانند: «کدام کشور x را ساخته است؟»، «محصول x ساخت کدام کشور است؟» و یا «کدام کشور تولید کننده x است؟». حال با استفاده از این الگوها و اجرای درخواست‌ها روی گراف دانش فارسی، سوال و پاسخ آن را تکمیل می‌کنیم. مثلا: «کدام کشور افسانه جومونگ را ساخته است؟» که پاسخ «کره جنوبی» از گراف دانش استخراج می‌شود. شایان ذکر است که الگوهای ساده، برآمده از پژوهش علی عبداللهی در آزمایشگاه پردازش زبان طبیعی دانشگاه امیرکبیر بوده و در این پژوهش مطابق آنچه در بالا بیان شد، با داده‌های درون گراف دانش تکمیل شدند (عبداللهی و همکاران، ۱۴۰۱).

نوع دیگر از الگوهای این مجموعه پرسش و پاسخ، الگوهای پرسش‌های پیچیده هستند که در این پژوهش ساخته و پرداخته شدند. در این نوع الگو، دو یا چند رابطه به زبان طبیعی بیان می‌شوند. مثلا پرسش‌های «نام پایتخت کشور سازنده محصول x کدام است؟» و یا «زبان رسمی کشوری که محصول x را ساخته‌است چیست؟» که به ترتیب دو جفت رابطه «پایتخت-کشور سازنده» و نیز «زبان رسمی-کشور سازنده» را در یک پرسش بیان می‌کنند. بسته به تعداد موجودیت‌های داخل نمونه گراف دانش، هر کدام از این الگوها قابل تکمیل و تکثیر هستند. اگر پرسش چندین جواب داشته باشد، تمامی جواب‌ها استخراج شده و در مجموعه داده قرار می‌گیرند. همچنین نکته قابل توجه دیگر این است که موجودیت ابتدایی سوال، و پاسخ‌های مربوطه باید به شکل شناسه عددی قابل تشخیص قرار بگیرند. زیرا برای تعبیه آن موجودیت‌ها به کمک مدل‌های تعبیه گراف دانش از شناسه عددی کمک می‌گیریم.

نمونه زیر دو سطر از مجموعه پرسش و پاسخ را نشان می‌دهد. در قسمت الف شناسه داخل پرانتز، موجودیت ابتدایی و مابقی، پاسخ(های) آن پرسش هستند. قسمت ب نیز تکمیل شده آن مثال آورده شده است.

---

1. Multi-hop
2. template

۱. الف. چه بازیگرانی در فیلم [۲۱۴۲۵۷۰] نقش داشته‌اند؟ ۲۱۴۲۹۵۵, ۲۱۴۲۶۸۴, ۲۱۴۲۶۱۳

ب. چه بازیگرانی در فیلم **ماه و پلنگ** نقش داشته‌اند؟ **علی شادمان، داریوش ارجمند، پوراندخت مهیمن**

۲. الف. زبان رسمی کشور سازنده [۲۱۵۶۰۵۹] چیست؟ ۱۷۵۲۸۴۷

ب. زبان رسمی کشور سازنده **بلک‌ادر** چیست؟ **زبان انگلیسی**

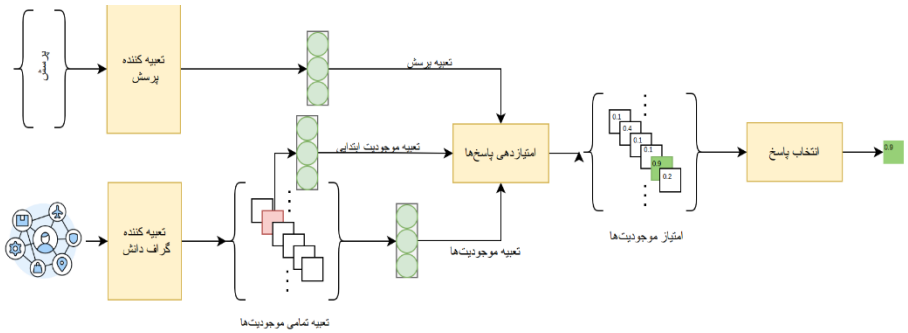
۴. مدل پیشنهادی

۴. ۱. شرح معماری سامانه پرسش و پاسخ

همانطور که در مقدمه نیز اشاره شد، معماری مد نظر در این پژوهش به تعبیه گراف و پرسش می‌پردازد که اساس این پژوهش برگرفته از معماری پیشنهادی مقاله تعبیه گراف دانش و پرسش و پاسخ است (Saxena et al., 2020). این سامانه که به منظور پاسخ به پرسش‌های چندپرشی در نظر گرفته شده است، فرض می‌کند با ساختن تعبیه مناسب و متناسب از پرسش مطرح شده، می‌تواند آن را به عنوان یالی فرضی برای رسیدن از موجودیت ابتدایی به موجودیت(های) انتهایی که همان پاسخ به پرسش است، برسد. با این فرض، حتی اگر موجودیت «سعدی» یال مشترکی با «حافظ» نداشته باشد اما با طی چندین یال به یکدیگر مسیر مشترکی داشته باشند، پرسش «شاعری که در محل تولد سعدی و در قرن هشتم می‌زیسته است؟» می‌تواند با تعبیه مناسب پرسش، به «حافظ» به عنوان پاسخ رسید. به‌طور خلاصه، به جای طی کردن چندین یال برای رسیدن به پاسخ، آن مسیر در درون تعبیه پرسش به‌صورت ضمنی قرار گرفته است.

شمای کلی این معماری در شکل ۴ آورده شده است. همانطور که مشاهده می‌شود، درون این معماری سه بخش اصلی وجود دارد که عبارت‌اند از تعبیه گراف دانش فارسی، تعبیه پرسش‌ها و انتخاب (تعبیه) پاسخ‌های مناسب از بین موجودیت‌های گراف دانش.

شکل ۴- معماری سامانه پرسش و پاسخ مبتنی بر گراف دانش



۴. ۲. مدل تعبیه گراف دانش

در این بخش با استفاده از مدل‌های مختلف از جمله کامپلکس، دیستمالت و غیره، از تمامی موجودیت‌ها تعبیه مناسب ساخته می‌شوند. این تعبیه‌ها بر اساس الگوها و ویژگی‌های نهفته در ساختار و تعاملات گراف دانش تهیه می‌شوند. به عبارتی به هر شناسه موجودیت[1] یک نمایش برداری نسبت داده می‌شود که بیانگر ویژگی‌های آن موجودیت در کل گراف دانش است.

تعبیه پرسش‌های زبانی

همانطور که گفته شد، پرسش مطرح شده در نقش یک یال فرضی عمل می‌کند که بر روی موجودیت ابتدایی اعمال می‌شود. بدین منظور در این قسمت از معماری، این ماژول وظیفه دارد پرسش مطرح شده را به فضای مشترکی با تعبیه‌های ساخته شده از گراف دانش ببرد. در گام اول پرسش با استفاده از مدل زبانی از پیش آموزش دیده شده، تعبیه شده تا درک مناسبی از صورت سوال بدست آید و سپس توسط چند لایه شبکه عصبی، به فضای هم‌اندازه با تعبیه موجودیت‌ها برده می‌شود.

در این پژوهش از مدل زبانی پیش آموزش دیده شده XLM-RoBERTa که مدل چندزبانه برت با آموزش بر روی ۲.۵ ترابایت داده متنی از ۱۰۰ زبان از جمله زبان فارسی است (Liu et al., 2019)، استفاده شده است. این قسمت تعبیه پرسش مدل زبانی را دریافت کرده و با وزن‌هایی که طی فرآیند آموزش تنظیم می‌کند، یاد می‌گیرد به فضای مشترک و هم‌اندازه

1. Entity ID

با فضای تعبیه موجودیت‌ها ببرد. در این بخش از شبکه‌های عصبی چندلایه[1] و لایه‌های حذف کننده تصادفی[2] برای تنظیم[3] استفاده می‌شود.

۳.۴. مدل واسط تعبیه پرسش و تعبیه گراف دانش

در آخرین بخش از این معماری، یک ماژول واسط وظیفه اتصال دو ماژول تعبیه گراف دانش و تعبیه پرسش را دارد و بدین وسیله تابع هزینه‌ای[4] را تعریف می‌کند که سیستم به‌طور کلی سعی می‌کند با استفاده از آن وزن‌های شبکه‌های خود را تنظیم کند. این ماژول تعبیه موجودیت ابتدایی که مشخصاً برای هر داده در مجموعه داده پرسش و پاسخ آورده شده است را با مدل تعبیه گراف دانش بدست می‌آورد. از سوی دیگر پرسش را واحدسازی[5] کرده و از شبکه عصبی چند لایه به همراه لایه حذف کننده تصادفی عبور می‌دهد تا به فضای متناسب با اندازه تعبیه موجودیت‌ها برسد. سپس با استفاده از تابع امتیازدهی مدل تعبیه گراف دانش(Φ)، دو تعبیه به دست آمده را به عنوان موجودیت ابتدایی(e_h) و رابطه(e_q) در نظر گرفته و به پیش‌بینی موجودیت‌هایی می‌پردازد که بیشترین امتیاز را به عنوان موجودیت انتهایی این سه‌گانه خواهند داشت. این موجودیت‌ها همان پاسخ پرسش(e_{ans}) هستند که در رابطه (۱) نمایش داده شده است.

$$e_{ans}=\text{argmax}\ \Phi(e_h, e_q, e_a);\ \text{for a} \in \mathcal{E} \qquad (۱)$$

تابع هزینه این مدل به‌صورت رابطه (۲) بیان می‌شود:

$$\text{Loss} = \text{KLDivLoss}(\log_{\text{-softmax}}(\text{scores}), \text{normalize}(\text{targets})) \qquad (۲)$$

که در آن از تابع هزینه واگرایی کولبک-لایبلر[6] بین امتیاز پاسخ‌های پیش‌بینی‌شده و پاسخ‌های اصلی استفاده شده است[7].

1. multilayer perceptron(MLP)
2. dropout
3. regularization
4. loss function
5. tokenize
6. Kullback-Leibler Divergence
7. تابع هزینه واگرایی کولبک-لایبلر میزان واگرایی دو توزیع احتمالاتی نسبت به یکدیگر را مشخص می‌کند.

۵. زمایش‌ها، نتایج و ارزیابی

۵. ۱. دادگان آموزش و ارزیابی

با توجه به حجم بالای گراف دانش فارسی، برای آموزش و ارزیابی بهتر مدل‌های تعبیه گراف دانش و با توجه به محدودیت منابع، نمونه‌گیری گراف دانش با هدف افزایش تراکم و تسکین تنک بودن گراف دانش انجام پذیرفته است. ساختار این نمونه‌گیری ۱۰۰-۱۰-۱۰ است، به این معنی که تمامی موجودیت‌های این نمونه حداقل ۱۰ همسایه دارند و همچنین هر یال در این نمونه، حداقل ۱۰۰ بار در کل گراف دانش تکرار شده است. شایان ذکر است که نمونه‌های مختلفی برداشته و کیفیت آنها ارزیابی شد که بهترین نمونه که در عین کوچک بودن، کیفیت مناسبی برای تعبیه اجزای گراف دانش مهیا کند این نمونه بود.

برای دادگان پرسش و پاسخ جهت ارزیابی، ۱۳۷۰ الگو برای ۶۰ رابطه ساده و ۶۲ الگو برای ۶ رابطه پیچیده ساخته شده است. شایان ذکر است در مجموعه پرسش و پاسخ تنها موجودیت‌هایی به کار رفته‌است که در نمونه گراف دانش موجود باشند، تا بتوان ارزیابی عادلانه و صحیحی از کیفیت مدل پیشنهادی ارائه داد. جدول ۲ تعداد پرسش‌ها این مجموعه داده را بیان می‌کند.

جدول ۲- مجموعه داده پرسش و پاسخ بر روی گراف دانش فارسی

نوع سوال	سوالات ساده	سوالات پیچیده
تعداد تکرار	۶۰۶۳۰	۱۳۷۸۷

۵. ۲. معیارهای ارزیابی

ارزیابی سامانه پرسش و پاسخ با استفاده از تعبیه گراف دانش در دو گام انجام شد.

- ارزیابی کیفیت تعبیه گراف دانش که براساس معیارهای زیر صورت گرفته است:
  - رتبه‌بندی وارون متوسط[1]
  - نرخ برخورد رتبه[2]
- ارزیابی کیفیت پاسخ‌گویی به پرسش‌ها که براساس معیارهای زیر صورت گرفته است:
  - دقت[3]
  - نرخ برخورد رتبه

1. Mean Reciprocal Rank(MRR)
2. Hit at K'th Rank(H@K)
3. accuracy

۳.۵. نتایج

در این بخش گراف دانش فارسی بر روی مدل‌های مختلفی آموزش دیده شده است. نتایج این ارزیابی در جدول ۳ نشان‌داده شده است. با توجه به نتایج موجود در جدول ۳ و نیز سادگی مدل، مدل کامپلکس برای قرارگیری در ماژول تعبیه سامانه پرسش و پاسخ انتخاب شد.

تعبیه‌های ساخته شده توسط کامپلکس به فضایی با ۲۵۶ بعد برده می‌شود. کیفیت این تعبیه‌ها به‌صورتی عینی و از طریق روش‌های بررسی چند نزدیک‌ترین همسایه[1] بررسی شد. آنچه مشاهده شد این بود که تعبیه‌هایی که بیشترین شباهت[2] را به‌هم داشتند، موجودیت‌هایی هستند که به‌صورت منطقی نیز به یکدیگر مربوط هستند و اطمینان حاصل شد بردارهای تعبیه ساخته شده، توانایی منتقل کردن دانش نهفته در گراف دانش را به‌خوبی در خود دارند.

جدول ۳- نتایج ارزیابی تعبیه‌های گراف دانش

مجموعه داده گراف دانش فارسی (۱۰۰-۱۰-۱۰)						
مدل		کامپلکس	دیس‌مالت	کانوای	ترنس‌ای	روتیت‌ای
صحت‌سنجی	MRR	۰.۰۳	۰.۰۱	۰.۰۴	۰.۰۱	۰.۰۲
	H@1	۰.۰۱	۰.۰۰۱	۰.۰۳	۰.۰۰۴	۰.۰۱۲
	H@10	۰.۰۵	۰.۰۳	۰.۰۶	۰.۰۳	۰.۰۳
	H@100	۰.۱	۰.۰۷	۰.۱۵	۰.۰۸	۰.۰۷
	H@1000	۰.۱۵	۰.۱۸	۰.۲۶	۰.۱۸	۰.۱۲
تست	MRR	۰.۱	۰.۰۱	۰.۱	۰.۰۴	۰.۰۸
	H@1	۰.۰۶	۰.۰۱	۰.۰۶	۰.۰۳	۰.۰۵
	H@10	۰.۱۷	۰.۱۰	۰.۱۸	۰.۰۷	۰.۱۵
	H@100	۰.۳۵	۰.۰۵	۰.۳۵	۰.۱۵	۰.۲۷
	H@1000	۰.۴۵	۰.۱۱	۰.۴۵	۰.۳	۰.۳۸

1. Kth Nearest Neighbor(KNN

2. در این جا از شباهت کسینوسی استفاده شد.

در بخش دوم ارزیابی به شرح آزمایش‌ها و نتایج پیرامون آموزش و ارزیابی سامانه پرسش و پاسخ با استفاده از گراف دانش فارسی می‌پردازیم. معماری انتخاب شده این پژوهش با استفاده از مدل تعبیه گراف دانش که در گام قبل به خوبی آموزش دید، هم اکنون بر روی مجموعه داده پرسش و پاسخ گراف دانش فارسی آموزش داده می‌شود. در معماری به کار رفته، از یک لایه متراکم[1] به اندازه ۲۵۶×۷۶۸ و نیز لایه حذف کننده تصادفی با ضریب فراموشی ۰.۱ برای تعبیه پرسش‌های زبان طبیعی به فضای مشترک با تعبیه گراف دانش استفاده شده است. تعبیه‌های گراف دانش نیز از یک نرمال‌ساز دسته‌ای[2] عبور می‌کند. در این معماری با ۱۰ ایپاک و هر دسته[3] ۲۵۶ تایی از سوالات تحت آموزش قرار گرفت. ادامه فرآیند آموزش مدل تا جایی رقم خورد که با ارزیابی مدل بر روی داده‌های اعتبارسنجی در پایان هر ۵ ایپاک، معیار دقت کمتر نشود.

نتایج کیفیت این سامانه در جدول ۴ آورده شده است. کیفیت سامانه در سه حالت، شامل فقط پرسش‌های پیچیده (پرسش‌های چندپرشی)، فقط پرسش‌های ساده و حالت کلی، شامل هر دو نوع پرسش پیچیده و ساده بررسی شدند. همان‌طور که مشاهده می‌شود مدل پیشنهادی توانسته است به دقت خیلی خوبی در پرسش‌های ساده و دقت قابل قبولی در پرسش‌های پیچیده دست یابد.

جدول ۴- نتایج ارزیابی سامانه پرسش و پاسخ گراف دانش فارسی

	مجموعه داده پرسش و پاسخ					
	پرسش‌های پیچیده		پرسش‌های پیچیده و ساده		پرسش‌های ساده	
نتایج	دقت	*H@10*	دقت	*H@10*	دقت	*H@10*
	۰.۵۴	۰.۸۲	۰.۸۵	۰.۹۴	۰.۹۲	۰.۹۶

در ادامه چندین پرسش منتخب که سامانه به درستی پاسخ داده است انتخاب و آورده شده است.

۱) **پرسش:** زبان رسمی کشوری که کشور سازنده مأموران مخفی (مجموعه تلویزیونی) است چیست؟
پاسخ سامانه: زبان انگلیسی(کشور سازنده این مجموعه انگلستان می‌باشد)

1. dense
2. batch normalization
3. batch

۲) **پرسش:** پایتخت کشوری که کشور سازنده دردسرهای عظیم است چیست؟
پاسخ سامانه: تهران (کشور سازنده این مجموعه ایران می‌باشد)

۳) **پرسش:** نام بچه عبدالمطلب چیست؟
پاسخ سامانه: ابوطالب

۴) **پرسش:** ورزش مورد علاقه هلن استفنز
پاسخ سامانه: دو و میدانی

۶. جمع‌بندی و نتیجه‌گیری

یکی از کاربردهای گراف دانش، مسئله پاسخ‌گویی به پرسش‌هایی است که جوابشان در دل ساختار گراف قرار گرفته است. این پرسش‌ها می‌توانند به‌صورت درخواست‌های منطقی مطرح و پاسخ‌گویی شوند. اما با توجه به پیشرفت‌هایی که در زمینه درک زبان طبیعی و مدل‌های زبانی بزرگ صورت گرفته است، می‌توان پا را فراتر از منطق گزاره‌ای گذاشت. در این پژوهش سامانه‌ای برای تعبیه گراف دانش و نیز پاسخ‌گویی به پرسش‌های زبانی طبیعی با استفاده از این مدل‌ها و نیز مدل‌های زبانی مبتنی بر ترانسفورمر پیشنهاد شد که بتواند به سوالات ساده و حتی پیچیده‌تر جواب بدهد. ابتدا نمونه‌های مختلفی از گراف دانش فارسی فارس-ویکی-کی‌جی استخراج شد تا فرآیند آزمایش و ارزیابی را تسهیل و تسریع ببخشد. سپس مدل‌های مختلفی برای تعبیه این گراف دانش استفاده و ارزیابی شدند. مجموعه‌داده‌ای برای پرسش‌های ساده و پیچیده برای پرسش‌ها به زبان طبیعی و پاسخ آنها از درون گراف دانش ارائه شد. در نهایت، معماری پیشنهادی آموزش داده شده و با معیارهای مختلف ارزیابی شد. به عنوان اولین پژوهش بر روی تعبیه گراف دانش فارس-ویکی-کی‌جی، آزمایش‌ها و ارزیابی‌های انجام شده نشان از قدرت و انعطاف پذیری بالای این معماری در پاسخ به پرسش‌های چندپرشی داشت. مدل کامپلکس به عنوان مدل تعبیه گراف دانش، در کنار مدل زبانی چندزبانه XLM-RoBERTa، باعث شدند هم درک خوبی از موجودیت‌ها و روابط بین آنها در گراف دانش پیدا کنیم و همچنین درک مطلوبی از پرسش‌هایی که به زبان طبیعی مطرح می‌شوند بدست آوریم. کارهای آینده می‌توانند شامل تغییر و ایجاد نوآوری در معماری مدل و عمیق کردن لایه‌های آن، و نیز ارزیابی و تنظیم عملکرد مدل بر روی مجموعه‌داده‌هایی با پرسش‌های پیچیده‌تر، که نیازمند استنباط‌هایی فراتر از طی چندین یال گراف دانش است، باشد.

تقدیر و تشکر

این اثر تحت حمایت مادی بنیاد ملی علم ایران (INSF) برگرفته شده از طرح شماره «۴۰۳۱۸۱۳» انجام شده است.

منابع

عبداللهی، علی. (۱۴۰۱). پیاده‌سازی سیستم یکپارچه گراف دانش فارسی به همراه هستان‌شناسی و پرسش و پاسخ [پایان‌نامه کارشناسی ارشد]. دانشگاه صنعتی امیرکبیر.

Asgari-Bidhendi, M., Hadian, A., & Minaei, B. (2019). FarsBase: The Persian knowledge graph. *Semantic Web, 10*(5), 1–15.

Bollacker, K. D., Evans, C., Paritosh, P. K., Sturge, T., & Taylor, J. (2008). Freebase: A collaboratively created graph database for structuring human knowledge. *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, 1247–1250. https://doi.org/10.1145/1376616.1376746

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. *Advances in Neural Information Processing Systems, 26*, 2787–2795.

Choudhary, S., Luthra, T., Mittal, A., & Singh, R. (2021). A survey of knowledge graph embedding and their applications. *arXiv preprint arXiv:2107.07842*.

Huang, X., Zhang, J., Li, D., & Li, P. (2019). Knowledge graph embedding based question answering. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 105–113. https://doi.org/10.1145/3289600.3290956

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., ... & Bizer, C. (2015). DBpedia—A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web, 6*(2), 167–195. https://doi.org/10.3233/SW-140134

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nickel, M., Tresp, V., & Kriegel, H. P. (2011). A three-way model for collective learning on multi-relational data. *Proceedings of the 28th International Conference on Machine Learning*, 809–816.

Rossi, A., Barbosa, D., Firmani, D., Matinata, A., & Merialdo, P. (2021). Knowledge graph embedding for link prediction: A comparative analysis. *ACM Transactions on Knowledge Discovery from Data, 15*(2), 1–49. https://doi.org/10.1145/3441454

Ren, H., Hu, W., & Leskovec, J. (2020). Query2box: Reasoning over knowledge graphs in vector space using box embeddings. *International Conference on Learning Representations*.

Saxena, A., Tripathi, A., & Talukdar, P. (2020). Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. *Proceedings of the 58th Annual Meeting of the Association for*

Computational Linguistics, 4498–4507. https://doi.org/10.18653/v1/2020.acl-main.412

Shirmardi, F., Hosseini, S. M. H., & Momtazi, S. (2021). Farswikikg: An automatically constructed knowledge graph for Persian. *International Journal of Web Research, 4*(2), 25–30.\

Sun, H., Bedrax-Weiss, T., & Cohen, W. (2019). PullNet: Open domain question answering with iterative retrieval on knowledge bases and text. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2380–2390. https://doi.org/10.18653/v1/D19-1242

Sun, H., Dhingra, B., Zaheer, M., Mazaitis, K., Salakhutdinov, R., & Cohen, W. (2018). Open domain question answering using early fusion of knowledge bases and text. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4231–4242. https://doi.org/10.18653/v1/D18-1455

Sun, Z., Deng, Z. H., Nie, J. Y., & Tang, J. (2019). RotatE: Knowledge graph embedding by relational rotation in complex space. *International Conference on Learning Representations*.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., & Bouchard, G. (2016). Complex embeddings for simple link prediction. *Proceedings of the 33rd International Conference on Machine Learning*, 2071–2080.

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A free collaborative knowledge base. *Communications of the ACM, 57*(10), 78–85. https://doi.org/10.1145/2629489

Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering, 29*(12), 2724–2743. https://doi.org/10.1109/TKDE.2017.2754499

Yang, B., Yih, W. T., He, X., Gao, J., & Deng, L. (2015). Embedding entities and relations for learning and inference in knowledge bases. *International Conference on Learning Representations*.

استناد به این مقاله: اسد، علی و ممتازی، سعیده. (۱۴۰۳). تعبیه گراف دانش به منظور بهبود سامانه‌های پرسش و پاسخ فارسی. *زبان و زبان‌شناسی*، ۲۰(۴۰)، ۱۳۱- ۱۵۲.

doi: 10.30465/LSI.2025.50863.1791