Intelligent Data Analysis 1 (2023) 1–5
IOS Press

Farspredict: A Benchmark Dataset for Link Prediction

Najmeh Torabian [a], Behrouz Minaei-Bidgoli [b,*] and Mohsen Jahanshahi [a]
[a] *Department of Computer Engineering, Central Tehran Branch, Islamic Azad University, Tehran, Iran*
E-mails: najmeh.torabian@gmail.com, mjahanshahi@iauctb.ac.ir
[b] *Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran*
E-mail: b_minaei@iust.ac.ir

Abstract. Link prediction using knowledge graph embedding (KGE) is a popular method for completing knowledge graphs. Moreover, training KGEs on non-English knowledge graphs can enhance knowledge extraction and reasoning within the context of these languages. However, several challenges in non-English KGEs hinder the learning of low-dimensional representations for a knowledge graph's entities and relations. This paper proposes "Farspredict," a Persian knowledge graph based on Farsbase, the most comprehensive Persian knowledge graph. It also explains how the knowledge graph structure affects link prediction accuracy in KGE. To evaluate Farspredict, we implemented popular KGE models on it and compared the results with those of Freebase. After analyzing the results, we carried out some optimizations on the knowledge graph to improve its functionality in KGE, resulting in a new Persian knowledge graph. The implementation results of KGE models on Farspredict outperformed Freebase in many cases. Lastly, we discuss possible improvements to enhance the quality of Farspredict and the extent to which it improves.

Keywords: Link prediction, Knowledge graph embedding, and benchmark knowledge graph

1. Introduction

Knowledge graphs have received much attention in recent years due to their applications, which offer significant economic benefits. A knowledge graph contains the knowledge obtained from various sources, including texts and tables. It has numerous applications in natural language processing and has been investigated as a potential reasoning source for explainable artificial intelligence.

Although the impact of creating knowledge graphs in non-English languages has been explored recently, little attention has been paid to preparing a suitable knowledge graph for use in the link prediction field.

At the same time, one of the main reasons why significant progress has not yet been made in Persian reasoning, recommendation systems, and other similar fields is the lack of a proper knowledge graph in these languages. Although some attempts have been made to construct a Persian knowledge graph, the most successful one is the Farsbase project. However, by applying Farsbase for link prediction through KGE models, we realized that it is too weak to be used for link prediction.

In the approach to state-of-the-art link prediction methods, we come to KGE methods. These methods were introduced with TransE, which falls into translational distance models. After TransE [6], the TransH [44], TransD [17], TransR [20] models, and many other methods that improved TransE and its

*Corresponding author. E-mail: b_minaei@iust.ac.ir.

other models. These methods convert triples into vectors and predict the new triple in the Freebase and WordNet graphs by considering the location of the embedded vectors. Almost another group of models known as semantic similarity emerged simultaneously with the translational distance models, including RESCAL [29], DistMult [46], HolE [30], and ComplEx [40]. Similar to the previous category, these models mostly use Freebase and WordNet datasets and apply matrix decomposition techniques to embed existing triple components and predict new triples. The latest and most widely used category of link prediction models in KGE methods is deep learning methods.

Despite the acknowledged importance of Persian datasets, they are rarely used experimentally for link prediction. Therefore, considering the practical application of these link prediction methods, we decided to survey the implementation of these models on non-English and particularly Persian knowledge graphs. Unfortunately, we found only a few studies in low-resource languages and no studies in Persian.

This study aims to develop a Persian knowledge graph for link prediction and use it through the KGE models for knowledge graph completion as a valuable knowledge reference and obtain embedded vectors to be applied in Persian applications. Therefore, this paper presents a set of steps to make changes in Farsbase and create a standard knowledge graph for link prediction. Based on this set of steps, it then describes the implementation of the KGE models on the prepared knowledge graph to determine the reliability of this dataset in this field.

The paper is structured as follows: Section 2 discusses the related works, while Section 3 details the proposed three-step standardization of our Persian knowledge graph and the implementation of the KGE models. Section 4 presents the experiments conducted on Farspredict and the results obtained, which are further analyzed in Section 5 for quality. The paper concludes in Section 6, along with a discussion on future research scope.

2. Related Works

Since knowledge graphs are created to understand and reason with human languages, human knowledge needs to be represented, stored, and extracted in a form that computers can process. Knowledge graphs have been developed as a knowledge base for entities and relations among these entities. However, they have a limitation in coping with most human languages.

On the other hand, KGE has provided extraordinary progress in representation over the years, but many still need to apply it to real datasets. The size, intricacy, and variety of these datasets significantly challenge their utilization.

Several partly human efforts have been invested in making KGs available across languages [19] [48]. However, even well-known KGs, including DBpedia [3], Wikidata [41], YAGO [39], Freebase [5], BabelNet [27], and Google knowledge graph [11], are most abundant in their English version. This lack of a multilingual knowledge graph limits the porting of natural language processing (NLP) tasks, such as link prediction, question answering, and recommendation systems, to different languages. However, several attempts have been made to construct knowledge graphs in other languages; examples are given below. The remarkable thing about these knowledge graphs is that there is no report indicating the use of these items in KGE and link prediction.

XLore [43] is an English-Chinese bilingual knowledge graph that enriches Chinese knowledge using online wikis. XLore's authors tried to cover semantic inconsistency between concepts, not equivalent cross-lingual entity linking problems. XLore2 [18] is an XLore extension established to eliminate the incompleteness of the dataset by adding facts via making cross-lingual knowledge linking, cross-lingual

property matching, and fine-grained type inference. Another non-English knowledge graph is Lynx [26]. The legal knowledge graph in the multilingual European knowledge graph is part of a big project for innovative compliance services. The Lynx project benefits from the technology-driven content product of KDictionaries, which developed quality lexicographic data in 50 languages. In addition, HOLINET [31] is a holistic KG for French, PolarisX [51] is an automated expansion KG, RezoJDM [24] in French, and VisualSem [1] is a visual multilingual knowledge graph.

Due to the increasing applications of knowledge in various aspects of human life, downstream requests are increasing, and the number of specific knowledge graphs is also growing. To clarify this article, we will explain some examples of these applications below.

The cultural heritage Chinese knowledge graph [13] provides updated information to its users and can grow over time. Another article in the Chinese language [21] collects information automatically with a crawler and uses machine learning tools to build the graph. Research by Marchand et al. [22] includes information on the cultural heritage of the province of Quebec in French. Bruns et al. [8] discuss efforts to preserve the culture of Nuremberg, while Tan et al. [38] present research on protecting the spread of book heritage in German. Arco [9] is another European knowledge graph in the Italian language, which is a product of a Cultural Heritage project that includes software, a documentation report, and other components.

Medicine Zheng and his colleagues created a particular purpose knowledge graph called TCMKG [53] to preserve traditional Chinese medicine. The COVID-19 pandemic inspired the construction of a knowledge graph based on the spatial distribution of the disease in the article by Yang et al. [47]. Other research, such as Feng et al. (2022) and Sakor et al. [15] and Sakor et al. [49] have also built knowledge graphs based on the distribution of the COVID-19 vaccine and disease. The content of these datasets is primarily obtained by establishing countries with research and development teams working on the vaccine and extracting entities and relations from them.

Other applications In addition to the examples mentioned above, there are other datasets in various applications. For instance, WeaKG-MF [4] is a knowledge graph that contains meteorological data collected in French. There is also a particular knowledge graph for solving electrical device problems through question answering in Chinese [23]. The European Union has a knowledge graph containing information from various data sources [36], and Mishra et al. [25] developed a specific knowledge graph for the tourist industry. Finally, Zehra et al. [52] provide an automatic query engine to find hidden relationships between financial documents.

In addition to the mentioned knowledge graphs, CAMS_KG [7] is a morpho-semantic knowledge graph that combines Ghwanmeh stemmer and MADAMIRA to support Arabic knowledge representation and information retrieval. Since Arabic is one of the languages closest to Persian, this study is essential. Besides, the last and most crucial non-English knowledge graph for us is Farsbase [2]. Farsbase is the first Persian knowledge graph for extracting knowledge from multiple sources. It obtains a rich knowledge graph from popular sources, including Wikipedia and online sites. Although this excellent knowledge graph is the primary source of the present research, it has some problems when used in the link prediction field.

Although many projects have been done to create non-English or multi-language knowledge graphs, no reliable Persian knowledge graph was found for link prediction. To the best of our knowledge, the literature has yet to discuss a Persian knowledge graph using knowledge graph representation and link prediction using the KGE models.

Although modern techniques exist for constructing knowledge graphs from primary sources such as text and tables, they often contain incomplete information. To extract knowledge from them, methods such as link prediction are needed to complete the graph.

The KGE models are a way to embed a knowledge graph in a vector space while retaining its main properties. These models are divided into three main categories, some of which are briefly discussed below. In these models, a triple is proposed to be added to the knowledge graph to complete the dataset through link prediction techniques, such as negative sampling [50].

Translational distance models are the first group of embedding models, in which each of the three elements (head, relation, and tail) is considered as a vector in the embedded space according to the chosen specific model, and the space of the elements can be shared or separate [42]. The first model in this category is TransE [6], which has achieved remarkable success. The scoring function used in this model is $h + r \approx t$. The main disadvantage of this model is that it covers only 1-to-1 triples. Other models include TransH [44], TransR [20], TransD [17], and several other models that try to solve this problem and improve the efficiency of TransE results and its other models.

Semantic matching models are based on semantic similarity matching and the relations between entities through the scoring function. Scoring functions in this model are essential in identifying new triples and introducing them to the knowledge graph. The first model in this set is RESCAL [29], in which entities are considered vectors, and the relations between these entities are matrix. Other models in this category try to improve the performance of the RESCAL model, such as DistMult [46], HolE [30], and ComplEx [40]. In these models, relational data is modeled as a three-way tensor X of size n×n×m, where n is the number of entities, and m is the number of relations. This knowledge representation model limits the scope of knowledge graph completion within the knowledge graph and cannot predict new knowledge beyond the knowledge graph.

Neural network models use deep neural network tools to learn the ternary model in a knowledge graph. Then, through the exploratory structure, propose new triples and complete the knowledge graph. One of the first methods to embed the knowledge graph based on deep learning is the ProjE model [34], in which entities and relations between them are embedded seamlessly and cohesively. This model emphasizes reducing the number of parameters, leading to less time complexity and less computation. Other models in this category include ConvKB [28], ConvE [10], RotatE [37], SACN [32], and many more. The deep neural network models have shown good predictive performance, even if they are more expensive and time-consuming compared to other categories.

In this research, Farsbase is considered a primary dataset on which the KGE models are implemented. The results of an empirical study showed that Farsbase has some inconsistencies to be used in link prediction and has brought the idea of making a new knowledge graph for KGE models-based link prediction.

3. Persian knowledge graph construction

Knowledge graphs are valuable resources that provide the possibility of extracting knowledge from textual sources. They are also used as a data reference for many technologies, including question-answering systems, recommender systems, and knowledge management. Therefore, this type of data reference is necessary for every language. Farsbase, as the first Persian knowledge graph, is a rich source, but it could be more effective for use in link prediction and completing the knowledge graph. Its shortcomings cause knowledge extraction to be disturbed. The process used to create the new knowledge graph is discussed in this section.

Two hypotheses have been proposed in studies on the structure of graph knowledge and knowledge graph embedding through multiple models. The first hypothesis is that creating a standard knowledge graph in the Persian language can yield link prediction results similar to those of KGE models. The second hypothesis is that deep learning methods have shown good performance in knowledge graph embeddings, and by implementing deep learning models on the obtained Persian knowledge graph, better accuracy can be achieved in link prediction compared to other methods.

To implement KGE models on the Persian dataset and evaluate its performance, we used the OpenKE framework that contains various graph knowledge embedding models [16]. At the time of writing this article, the source code of this framework is open and includes nine valid KGE models. The framework has been used in research and implementation on two datasets, Freebase [5], and WordNet [14]. We implemented the models from this framework on Farsbase graph knowledge. However, the results were significantly weaker than those reported in the articles related to the embedding models of the knowledge graph. To address these shortcomings, we created a new Persian knowledge graph called Farspredict for link prediction based on the Farsbase dataset. In the following section, we will describe the specifications of Farsbase and introduce Farspredict.

3.1. Farsbase Specifications

To obtain embedded vectors of the knowledge graph for link prediction, we need to implement the KGE models on our Persian knowledge graph. As Farsbase is the first Persian knowledge graph, we used this dataset to obtain the embedded knowledge graph. However, it was necessary to prepare a version of the knowledge graph to do this. The first change was related to the ontological structure of Farsbase. We needed triples with two entities and a relation between them, not properties or anything else. The specifications of this version are listed in Table 1.

Table 1
Specifications of a version of the Farsbase Knowledge Graph

Dataset	# relations	# entities	# triples
Farsbase	7378	541927	2398999

Despite being a rich Persian knowledge graph compiled from unsupervised and unstructured texts, many of the triples in Farsbase do not represent factual information and lack inferential value. Moreover, a substantial portion of the relations in Farsbase has only a few associated facts. Specifically, 71% of relations have fewer than 100 facts, and many of them are limited to only one fact per relation type. This results in a knowledge graph that is both large and sparse.

The next step in providing a representation of the Persian knowledge graph and comparing the results with standard datasets is to implement OpenKE models on the proposed version of Farsbase. To implement KGE models, we divided this version of the dataset extracted from Farsbase into three sections: training, testing, and validation, and then applied the KGE models to them. In this phase, we allocated 10% of the dataset for the validation dataset, 20% for the testing dataset, and 70% for the training dataset. We applied six translational distance models, including TransE, TransH, TransR, TransD, TransG [45], and TransM [12], as well as four models of semantic matching models, namely RESCAL, HolE, ComplEx, and DistMult. The output is based on two metrics, Hits@10 and Mean Rank achieved. The mean rank is the average rank for all predicted triples within each model $((1 + n))/2$, and the proportion of testing triples whose ranks are not larger than 10 is HITS@10. This is called the "Raw" setting. When

we filter out the corrupted triples in the training, validation, and testing datasets, it is called the “Filter” setting. If a corrupted triple exists in the knowledge graph, ranking it before the original triple is acceptable. A higher HITS@10 and a lower mean rank mean excellent performance.

The results obtained from implementing the models on Farsbase are significantly weaker compared to the same models on the Freebase and WordNet datasets. This can be attributed to the fact that the degree of nodes in the Farsbase graph is highly uneven, with many entities and relation types appearing in only one triple, while only a small number of entities have a degree greater than 30,000. As explained in the "Related Works" (Section 2), the Farsbase dataset was prepared in RDF format and is a rich source of the first Persian knowledge. However, it is not yet suitable for implementing link prediction applications and can only be used as a reliable data source.

3.2. Farspredict

To validate Farsbase for link prediction, we needed to make changes to the dataset to make it suitable for use with KGE models. The process we used to create a standardized knowledge graph called Farspredict was simple, rapid, and reliable. The process involved the following steps.

These modifications are made based on KGE models to enable the use of knowledge graphs in link prediction. The removal of images or hyperlinks is done for this purpose, indicating that the presence of an image or hyperlink does not comply with the properties of the knowledge graph.

The standardization process for Farspredict consists of three parts.

The first part involves modifying the content of the triples. Some entities are in English or Arabic, while others contain non-textual items. To prepare the content in this section, two correction steps are presented. In the first step, entities in collected triples, images, or URLs are removed. In the second step, non-Persian entities and relation types are removed. For example, "isA" and "isRelatedTo" in the knowledge graph are correct as a triple in the knowledge graph, but they are omitted because they are non-Persian. The same procedure is applied to entities in non-Persian languages. After the implementation of this section, the number of triples decreased to 1.5 million.

Part Two Graph Structural Modification The existing graph is disconnected and has single nodes up to this point. We follow the following two steps to optimize the structure. Step 1: Evaluate the graph to find the number of unrelated subgraphs. At the end of this evaluation, it was found that this graph consists of 48 components, with the largest subgraph accounting for 99.87% of all triples in the graph. Step 2: We remove small subgraphs and single nodes.

Part three Content Modification in the Knowledge Graph In addition to being irregular, the graph obtained from the previous sections shows that the number of triples of different relation types is widely varied. To regularize the graph and distribute the triples better in the graph, we present five optimization levels in this section, and the order of execution is significant. Level One: We calculate the number of triples of each relation type. Level Two: We remove less than 100 triples for each relation type. Level Three: We remove the triples containing the deleted relations. Level Four: We remove isolated nodes (incoherent components) due to removing relations. Level Five: We remove the entities and relations that no longer exist during graph standardization and triple elimination from the entities and relations repository.

These steps are shown in Algorithm 1 and Figure 1.

After following the standardization steps, the first version of the new dataset, Farspredict, was created.

However, the implementation results were still below expectations and weaker than the Freebase results. Further examination of the graph revealed that there were chains in some parts of the current

Algorithm 1 Knowledge graph standardization

Require: Knowledge graph Farsbase,dts, Farspredict
Require: head h, tail t, relation r, triple tp, ent[], rel[]

```
while tp in Farsbase do
    if t and h in ent [] then
        dts = dts ∪ tp
    end if
end while
Subgraphs = cgSpan [33](dts)
set FarsPredict = biggest subgraph in Subgraphs
Remove images and U RLs from ent[]
Remove non Persian r from rel[] and t, h from ent[]
Repeat lines 6 and 7
while r in rel[] do
    if tp(r) < 100 then
        Remove r from rel[]
        Remove tp(r)
    end if
end while
while e in ent[] do
    if factchecker[35] (e) < 5 then
        Remove e from ent[]
        Remove tp(e)
    end if
end while
Repeat lines 6 and 7
Return Farspredict
```

graph. To reduce graph chains, we removed entities with less than five connections (entities connected to less than five other entities) from the dataset. Finally, the specifications of the second version of the Farspredict knowledge graph were determined, as shown in Table 2.

Table 2
Specifications of the final version of the Farspredict

Dataset	# relations	# entities	# triples
Farspredict	392	107827	622287

4. Experiments and Results

In this section, we describe the experiments conducted on KGE models using a Persian knowledge graph to generate embedded vectors of triples. It is expected that after going through the steps of constructing the Persian knowledge graph in the previous section, significant progress will be made, and the prerequisites for link prediction will be provided.

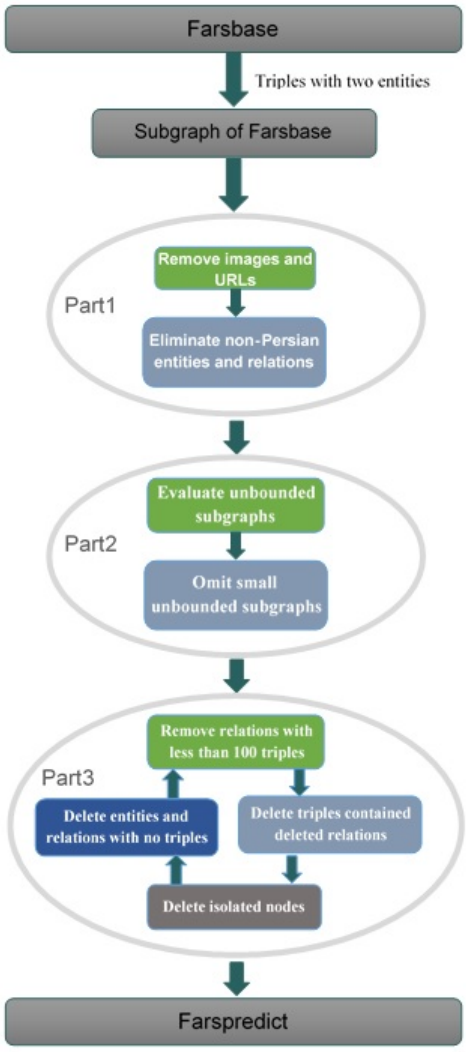

Fig. 1. Standardization Schema

We compared the knowledge graph created in Section 2 with the benchmark datasets. The ratio of the number of entities to the number of relation types in Freebase and Farspredict indicates that the new graph is sparse, and there are relation types with low frequency.

After implementing the KGE models on the Farspredict knowledge graph, the results were found to be weaker than expected, and even weaker than the Freebase results. Upon further examination of the graph, it was discovered that some parts of the graph contained chains. To address this issue, entities with fewer than five degrees (i.e., entities connected to fewer than five other entities) were removed from the dataset to reduce the chains in the graph.

Human evaluation First, we randomly select 1000 triples from Farspredict, remove the tails of these triples, and give the incomplete triples to three human experts with basic knowledge of knowledge graphs. The results obtained from the human evaluation are 95.1%, 93.4%, and 93% accurate. The high average of 93.83% demonstrates that the dataset triples are mostly understandable to humans.

Application by link prediction KGE models need to be implemented on the Farspredict knowledge graph to be examined. We used the OpenKE framework to access KGE models and implemented them on both Farspredict and Freebase datasets. The results are presented in Table 3.

Table 3
Execution of KGE models on the final version of Farspredict

	Mean Rank		Top10	
Models	Raw	Filtered	Raw	Filtered
TransE	3532.356	2980.389	0.318	0.374
TransH	4509.083	3958.75	0.291	0.338
TransD	4290.836	3732.386	0.294	0.343
DistMult	8176.77	7676.532	0.165	0.171
ComplEx	8228.179	7703.388	0.112	0.125
RESCAL	96559.422	96439.093	0.0004	0.0004
Analogy	9594.363	9035.328	0.21	0.223
Rotate	2327.429	1763.87	0.365	0.439
SimplE	7837.303	7338.044	0.166	0.174

As shown in Table 3, the results improved sufficiently, and the dataset results became closer to the standard dataset in the ratio of the number of relations to the number of triples. The Persian knowledge graph is still sparse despite general satisfaction with the results. We will use link prediction to eliminate the sparsity of the knowledge graph and complete it.

4.1. Results

Graph connectivity was effective, at least in our experiments, but the dataset's dispersion and heterogeneity, the graph's non-uniformity, the aggregation of the triples, and the degree of the node non-uniformity challenge any analysis and inference in the knowledge graphs.

While the only significant changes made were removing relations with less than 100 triples and entities with less than five connections, the Raw MRR in the TransE model differed by approximately 1311. The implementation results of the other models were also similar to TransE. Two hypotheses could account for this difference. Hypothesis 1 suggests that the presence of entities that are rarely used in Wikipedia reduces the likelihood of selecting a new valid triple.

Some entities cause chain formation in the graph, making the knowledge graph and inferring and technically predicting from this graph a challenging task. Since the points they earn in the scoring function are low, they will not play a role in predicting the triples and will not be selected. These changes are shown in Table 4 and Figure 2.

Table 4
Farsbase and Farspredict triple frequency for entities

Entity degree	Farsbase	Farspredict
$n \leqslant 5$	754432	65834
$5 < n \leqslant 50$	76233	2068
$50 < n \leqslant 500$	1273	183
$500 \leqslant n$	13	100

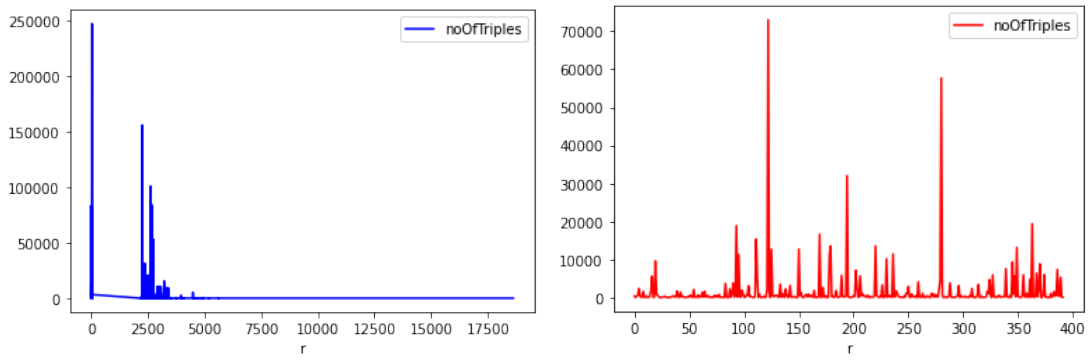

Fig. 2. Triple numbers based on relation types. a) Farsbase, b) Farspredict

Hypothesis 2: Relations with few triples are private ones used in a specific context that cannot be generalized to other cases. Including these types of triples can reduce link prediction accuracy. Setting a threshold to ignore these relations can help to address the issue of the graph's scattering and heterogeneity. These changes are shown in Table 5.

Table 5
Farsbase and Farspredict triples' frequency for relations

	Sum of triples in a range	
#triples	Farsbase	Farspredict
$n \leqslant 100$	15810	141
$100 < n \leqslant 1000$	436	174
$1000 < n \leqslant 10000$	139	61
$10000 \leqslant n$	45	16

Overall, compared to standard dataset results, the results presented in Table 3 show that the mean rank of Farspredict link prediction is higher than standard datasets. This happened while our final dataset was approximately equal to Freebase. This difference in value in criterion "Mean Rank" could be because Farspredict is still sparse, and the number of entities is 107827, while the number of FB15K's entities is 15951. Although the number of triplets is not very different, the number of entities is about 7.2 times larger.

5. Discussion

A knowledge graph completion is a valuable technique to ensure better knowledge extraction and inference. In this regard, link prediction through the KGE model is used for knowledge graph completion. Since the lack of using this tool in the Persian language can be seen, in this study, a knowledge graph was created for link prediction, and then its embedding was done in vector format with knowledge graph embedding models.

The results obtained in section 3 are consistent with the outputs of the graph knowledge embedding models on Freebase, supporting the first hypothesis that these models can be implemented on Farspredict in a similar manner. Additionally, the results reported in Table 3 demonstrate the effectiveness of the deep learning method, thus supporting the second hypothesis.

In this project, we conducted a comprehensive investigation of the factors and steps involved in transforming our dataset into a link prediction in the knowledge graph. We identified certain procedures for altering the graph structure that could enhance the quality of the proposed knowledge graph for link prediction. We observed that graph connectivity is a crucial factor that significantly reduces the mean rank of KGE models. Other factors that facilitated the mean rank factor were the removal of relations and entities with negligible dispersion. The positive correlation between reducing the number of relation types and the mean rank factor is evident in our study. Our study includes the most significant prospective analysis of link prediction on the Persian knowledge graph to date.

The embedded Persian knowledge graph and the suggested triples from link prediction can be utilized in various fields that require knowledge extraction, such as reasoning on knowledge graphs. Previous studies have reported on numerous multilingual knowledge graphs and KGE models, which served as motivation for our project in Persian.

Despite the final analysis demonstrating the advantages of the new Persian knowledge graph, certain shortcomings may be attributed to the sources used to extract the initial triples. These sources are mainly Persian Wikipedia pages whose contents have not been validated. Additionally, similar verbs and words may appear in various contexts, leading to potential ambiguity and errors in the extracted knowledge.

Consistent with previous studies, we observed that the number of triples in Farspredict is similar to that of Freebase, and the accuracy of link prediction on embedded vectors is close to the results of Freebase, and in some cases, even better.

Researchers working on knowledge graph completion may benefit from the proposed graph. We hope that conducting additional studies and projects in this field will lead to a more robust and complete Persian knowledge graph that can be used as a knowledge reference in artificial intelligence projects.

6. Conclusion

A Persian knowledge graph is essential in various fields, including link prediction. However, it was found that the only Persian knowledge graph available, Farsbase, cannot meet the needs of link prediction. It was concluded that the main problem with the graph structure is that it needs to be uniform, and the triples should be evenly distributed on the surface of the graph. Besides the graph structure, removing several extracted triples from Wikipedia whose components were in other languages, such as Arabic, and triples with two entities contributed to the structural problems in the dataset, ultimately weakening the link prediction results. Implementing the KGE models on the latest version of the knowledge graph shows that using KGE models for link prediction in other languages not only leads to excellent achievements but also pays attention to the quality of the knowledge graph, resulting in better outcomes than popular knowledge graphs. Researchers working on knowledge graph completion can benefit from the proposed graph. We hope that additional studies and projects in this field will result in a more robust and complete Persian knowledge graph to serve as a knowledge reference in artificial intelligence projects.

Various changes in the knowledge graph dataset were evaluated by implementing KGE models and using the mean rank and Top@10 factors as evaluation metrics. We hope that our findings will have an impact on link prediction in Persian knowledge graphs. In future work, we plan to refine the proposed KG by incorporating data from other sources.

Accessing the first version of Farspredict opens up opportunities for further research and implementations in this area. One potential future direction could involve establishing a relationship hierarchy based on Schema.org as a reference taxonomy. Additionally, combining Farspredict with other valid knowledge graphs and creating multilingual knowledge graphs could be valuable endeavors.

The knowledge graph can be applied to any knowledge retrieval task. However, it is unclear whether our approach can be extended to other knowledge graphs beyond those used in this study.

7. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] H. Alberts, N. Huang, Y. Deshpande, Y. Liu, K. Cho, C. Vania, and I. Calixto. Visualsem: a high-quality knowledge graph for vision and language. In Proceedings of the 1st Workshop on Multilingual Representation Learning, pages 138–152, 2021.

[2] M. Asgari-Bidhendi, A. Hadian, and B. Minaei-Bidgoli. Farsbase: The persian knowledge graph. Semantic Web, 10(6):1169–1196, 2019.

[3] 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, pages 722–735. Springer, 2007.

[4] N. Y. Ayadi, C. Faron, F. Michel, F. Gandon, and O. Corby. Weakg-mf: a knowledge graph of observational weather data. In The Semantic Web: ESWC 2022 Satellite Events, 2022.

[5] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pages 1247–1250, 2008.

[6] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In Neural Information Processing Systems (NIPS), pages 1–9, 2013.

[7] I. Bounhas, N. Soudani, and Y. Slimani. Building a morpho-semantic knowledge graph for arabic information retrieval. Information Processing & Management, 57(6):102124, 2020.

[8] O. Bruns, T. Tietz, M. B. Chaabane, M. Portz, F. Xiong, and H. Sack. The nuremberg address knowledge graph. In European Semantic Web Conference, pages 115–119. Springer, 2021.

[9] V. A. Carriero, A. Gangemi, M. L. Mancinelli, L. Marinucci, A. G. Nuzzolese, V. Presutti, and C. Veninata. Arco: The italian cultural heritage knowledge graph. In International Semantic Web Conference, pages 36–52. Springer, 2019.

[10] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel. Convolutional 2d knowledge graph embeddings. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.

[11] L. Ehrlinger and W. W¨oß. Towards a definition of knowledge graphs. SEMANTiCS (Posters, Demos, SuCCESS), 48(1-4):2, 2016.

[12] M. Fan, Q. Zhou, E. Chang, and F. Zheng. Transition-based knowledge graph embedding with relational mapping properties. In Proceedings of the 28th Pacific Asia conference on language, information and computing, pages 328–337, 2014.

[13] T. Fan and H. Wang. Research of chinese intangible cultural heritage knowledge graph construction and attribute value extraction with graph attention network. Information Processing & Management, 59(1):102753, 2022.

[14] C. Fellbaum. Wordnet. In Theory and applications of ontology: computer applications, pages 231–243. Springer, 2010.

[15] Y. Feng, N. Zhang, H. Xu, R. Wang, and Y. Zhang. Visual analysis of the national characteristics of the covid-19 vaccine based on knowledge graph. In International Conference of Pioneering Computer Scientists, Engineers and Educators, pages 262–272. Springer, 2022.

[16] X. Han, S. Cao, X. Lv, Y. Lin, Z. Liu, M. Sun, and J. Li. Openke: An open toolkit for knowledge embedding. In Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations, pages 139–144, 2018.

[17] G. Ji, S. He, L. Xu, K. Liu, and J. Zhao. Knowledge graph embedding via dynamic mapping matrix. In Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers), pages 687–696, 2015.

[18] H. Jin, C. Li, J. Zhang, L. Hou, J. Li, and P. Zhang. Xlore2: large-scale crosslingual knowledge graph construction and application. Data Intelligence, 1(1):77–98, 2019.

[19] G. A. Lakshen, V. Janev, and S. Vrane˘s. Challenges in quality assessment of arabic dbpedia. In Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, pages 1–4, 2018.

[20] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu. Learning entity and relation embeddings for knowledge graph completion. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 29, 2015.

[21] S. Liu, H. Yang, J. Li, and S. Kolmani˘c. Preliminary study on the knowledge graph construction of chinese ancient history and culture. Information, 11(4):186, 2020.
[22] E. Marchand, M. Gagnon, and A. Zouaq. Extraction of a knowledge graph from french cultural heritage documents. In ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, pages 23–35. Springer, 2020.
[23] F. Meng, S. Yang, J. Wang, L. Xia, and H. Liu. Creating knowledge graph of electric power equipment faults based on bert–bilstm–crf model. Journal of Electrical Engineering & Technology, pages 1–10, 2022.
[24] M. Mirzapour, W. Ragheb, M. J. Saeedizade, K. Cousot, H. Jacquenet, L. Carbon, and M. Lafourcade. Introducing rezojdm16k: a french knowledge graph dataset for link prediction. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 5163–5169, 2022.
[25] R. K. Mishra, H. Raj, S. Urolagin, J. A. A. Jothi, and N. Nawaz. Cluster-based knowledge graph and entity-relation representation on tourism economical sentiments. Applied Sciences, 12(16):8105, 2022.
[26] E. Montiel-Ponsoda and V. Rodr´ıguez-Doncel. Lynx: Building the legal knowledge graph for smart compliance services in multilingual europe. In 1st Workshop on Language Resources and Technologies for the Legal Knowledge Graph, page 19, 2018.
[27] R. Navigli and S. P. Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial intelligence, 193:217–250, 2012.
[28] 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. Semantic Web, 10(5):947–960, 2019.
[29] 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 International Conference on Machine Learning, pages 809–816, 2011.
[30] M. Nickel, L. Rosasco, and T. Poggio. Holographic embeddings of knowledge graphs. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 30, 2016.
[31] J.-P. Prost. Holinet: Holistic knowledge graph for french. In Journ´ees Jointes des Groupements de Recherche*Linguistique Informatique, Formelle et de Terrain*(LIFT) et
Traitement Automatique des Langues(TAL), pages 123–130. CNRS, 2022.
[32] 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 Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 3060–3067, 2019.
[33] Z. Shaul and S. Naaz. cgspan: Closed graph-based substructure pattern mining. In 2021 IEEE International Conference on Big Data (Big Data), pages 4989–4998. IEEE, 2021.
[34] B. Shi and T. Weninger. Proje: Embedding projection for knowledge graph completion. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 31, 2017.
[35] P. Shiralkar, A. Flammini, F. Menczer, and G. L. Ciampaglia. Finding streams in knowledge graphs to support fact checking. In 2017 IEEE International Conference on Data Mining (ICDM), pages 859–864. IEEE, 2017.
[36] A. Soylu, O. Corcho, B. Elvesæter, C. Badenes-Olmedo, F. Y. Mart´ınez, M. Kovacic, M. Posinkovic, I. Makgill, C. Taggart, E. Simperl, et al. Enhancing public procurement in the european union through constructing and exploiting an integrated knowledge graph. In International Semantic Web Conference, pages 430–446. Springer, 2020.
[37] Z. Sun, Z.-H. Deng, J.-Y. Nie, and J. Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. arXiv preprint arXiv:1902.10197, 2019.
[38] M. A. Tan, T. Tietz, O. Bruns, J. Oppenlaender, D. Dess´ı, and H. Sack. Ddb-kg: The german bibliographic heritage in a knowledge graph. In HistoInformatics@ JCDL, 2021.
[39] T. P. Tanon, G. Weikum, and F. Suchanek. Yago 4: A reason-able knowledge base. In European Semantic Web Conference, pages 583–596. Springer, 2020.
[40] T. Trouillon, J. Welbl, S. Riedel, E. Gaussier, and G. Bouchard. Complex embeddings for simple link prediction. In International Conference on Machine Learning, pages 2071–2080. PMLR, 2016.
[41] D. Vrande˘ci´c and M. Kr¨otzsch. Wikidata: a free collaborative knowledgebase. Communications of the ACM, 57(10):78–85, 2014.
[42] Q. Wang, Z. Mao, B. Wang, and L. Guo. Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering, 29(12):2724–2743, 2017.
[43] Z. Wang, J. Li, Z. Wang, S. Li, M. Li, D. Zhang, Y. Shi, Y. Liu, P. Zhang, and J. Tang. Xlore: A large-scale english-chinese bilingual knowledge graph. In International semantic web conference (Posters & Demos), volume 1035, pages 121–124, 2013.
[44] Z. Wang, J. Zhang, J. Feng, and Z. Chen. Knowledge graph embedding by translating on hyperplanes. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 28, 2014.
[45] H. Xiao, M. Huang, and X. Zhu. Transg: A generative model for knowledge graph embedding. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2316–2325, 2016.
[46] B. Yang, S. W.-t. Yih, X. He, J. Gao, and L. Deng. Embedding entities and relations for learning and inference in knowledge bases. In Proceedings of the International Conference on Learning Representations (ICLR) 2015, 2015.

[47] X. Yang, W. Li, Y. Chen, and Y. Guo. Construction of a covid-19 pandemic situation knowledge graph considering spatial relationships: A case study of guangzhou, china. ISPRS International Journal of Geo-Information, 11(11):561, 2022.
[48] P. Fafalios, A. Kritsotaki, and M. Doerr. The SeaLiT Ontology–An Extension of CIDOC-CRM for the Modeling and Integration of Maritime History Information. ACM Journal on Computing and Cultural Heritage, ACM New York, NY, 2023.
[49] A., Sakor, S., Jozashoori, E., Niazmand, A., Rivas, K., Bougiatiotis, F., Aisopos, E., Iglesias, P.D., Rohde, T., Padiya, A., Krithara, et al., 2023. Knowledge4covid-19: A semantic-based approach for constructing a covid-19 related knowledge graph from various sources and analyzing treatments' toxicities. Journal of Web Semantics 75, 100760.
[50] U., Bharambe, C., Narvekar, P., Andugula, 2023. Ontology and knowledge graphs for semantic analysis in natural language processing, in: Graph Learning and Network Science for Natural Language Processing. CRC Press, pp. 105–130.
[51] S. Yoo and O. Jeong. Automating the expansion of a knowledge graph. Expert Systems with Applications, 141:112965, 2020.
[52] S. Zehra, S. F. M. Mohsin, S. Wasi, S. I. Jami, M. S. Siddiqui, and M. K.-U.-R. R. Syed. Financial knowledge graph based financial report query system. IEEE Access, 9:69766–69782, 2021.
[53] Z. Zheng, Y. Liu, Y. Zhang, and C. Wen. Tcmkg: A deep learning based traditional chinese medicine knowledge graph platform. In 2020 IEEE International Conference on Knowledge Graph (ICKG), pages 560–564. IEEE, 2020.