FarsWikiKG:

an Automatically Constructed Knowledge Graph for Persian

Farhad Shirmardi, Mohammad Hadi Hosseini, Saeedeh Momtazi*
Computer Engineering Department
Amirkabir University of Technology
Tehran, Iran
shirmardi@aut.ac.ir,hadi.hosseini@aut.ac.ir,momtazi@aut.ac.ir

Received: 2021/10/16 Revised: 2021/12/05 Accepted: 2021/12/24

***Abstract*—We present FarsWikiKG, a Persian knowledge graph extracted from Wikipedia. Wikipedia infoboxes have been used as a valuable resource for building knowledge graphs in recent years. FarsWikiKG consists of more than 2 million entities, as well as 5.7 million facts about the entities. Using Wikidata, we constructed an ontology with more than 6000 classes representing entity types. As the second Persian knowledge graph, which has the ability of self-update, FarsWikiKG shows improvement on NLP tasks, especially question answering systems. Although FarsWikiKG is a dynamic knowledge graph, our evaluation shows a coverage of 90% on Persian Wikipedia pages. As Wikipedia information is constantly changing, a fixed knowledge graph can provide unstable data to the user. The proposed system, in addition to solving the problem of unstable data, reduces the need for experts to extract and construct knowledge graphs manually. Storing information in RDF as a standard method of storing knowledge graph information, FarsWikiKG allows NLP systems to run SPARQL queries on it.**

Keywords*— *Knowledge Graph; Wikipedia; RDF; Persian

1. INTRODUCTION

Recent developments on information retrieval systems like search engines, chatbots, and Question Answering (QA) systems, have increased the need for more and up-to-date data sources. Knowledge graph is used as a resource for storing this information in RDF format, which includes entities (e.g., people, locations) and their relations (e.g., birthDate, address).

With the publication of several large-scale datasets [1, 2] and the rapid development of deep learning techniques, great advances in QA systems have been made. In particular, recent years have seen the growth of open-domain text-based QA research, an important branch of QA on large resources such as Wikipedia and the Web [3]. Many Question answering systems like [4, 5, 6] use the knowledge graph as a resource for finding answers.

Using a knowledge graph as side information in recommender systems, caused improvement in the user experience. These systems use items and their attributes from knowledge graph to understand relations between items [7].

In this paper we present FarsWikiKG, a Persian dynamic knowledge graph that can be used in different NLP tasks. Considering Persian as a low-resource language, Farsbase [8] is the only available Persian knowledge graph and our work is the second knowledge graph developed for this language. The extraction and processing operations of Farsbase, however, have been done once and data stored in RDF format. As information in Wikipedia is updated constantly, there is a need for a system that automatically updates or adds the data. FarsWikiKG is building a larger knowledge graph compared to Farsbase and providing the possibility of self-updating on this resource to solve the issue of out of date information and limited entities.

FarsWikiKG was constructed mainly from Wikipedia infoboxes as a rich collection of entities and Wikidata as an ontology data source. This method is the same as the method used in YAGO [9] as one of the first automatically knowledge graph construction systems. Wikidata, as a great collection of knowledge, has more than 50M types for entities in FarsWikiKG.[1] Using these types, FarsWikiKG is able to construct an ontology tree.

The rest of this paper is organized as follows: In Section 2, we discuss related works on available knowledge graphs. Section 3 describes the FarsWikiKG construction framework. In Section 4, FarsWikiKG evaluation and statistics are reported. Finally, we present our conclusions and future works in Section 5.

2. RELATED WORKS

YAGO [9] was presented as a light-weight and extensible ontology with high coverage and quality with more than 1 million entities and 5 million facts. These facts have been automatically extracted from Wikipedia and unified with WordNet using heuristic methods. YAGO used hierarchical category pages in Wikipedia to extract facts instead of using infoboxes. Category pages are lists of articles that belong to a specific category. As WordNet provides a cleaner hierarchy of these concepts rather than Wikipedia, YAGO uses both sources to achieve high accuracy.

YAGO2 [10] was introduced as an extension of YAGO, by focusing on temporal and spatial knowledge. In addition to Wikipedia and WordNet, YAGO2 uses GeoNames to extract nearly 10 million entities and events, as well as about 80 million facts.

YAGO3 [11] uses Wikipedia in multiple languages and also English WordNet to extract multilingual information. This technique added 1 million new entities and 7 million new facts.

[1] the numbers given in the paper about Wikidata are valid as of Mar. 28, 2022

10.22133/ijwr.2022.337760.1112 20.1001.1.26454335.2021.4.2.4.2

YAGO4 [12] is the latest version of YAGO knowledge graph, which combines selection of Wikidata classes as lower-level classes and schema.org as top-level classes. This results in a consistent ontology with 2 billion type-consistent triples for 64 million entities.

DBpedia [13] extracted data from multiple language editions of Wikipedia. This project consists of over 1.86 billion facts about 13.7 million entities. English Wikipedia, along with 110 other languages, are the resources used to construct this knowledge base. DBpedia mapped infoboxes of 27 different languages of Wikipedia into a single ontology consisting of 320 classes and 1,650 properties.

Wikidata [14] is a free knowledge base, providing data in all languages which can be read and edited by humans and machines. The main purpose of creating Wikidata was to create a rich source of data so that the data could be used to create projects related to Wikipedia. Wikidata supports more than 280 different languages. The Wikidata community is very similar to Wikipedia in that users have the ability to register in Wikidata and then start modifying, creating, or deleting Wikidata pages. The Wikidata knowledge graph holds information in triples, except that Wikidata uses different concepts such as items and phrases to store information. Items are equivalent to entities in other knowledge graphs, and phrases are equivalent to key and value pairs. The amount of information in Wikidata is more than 94 million items.

Freebase [15] contains more than 125 million tuples, more than 4,000 types, and more than 7,000 properties. Freebase as a practical knowledge graph, collected information from multiple sources including Wikipedia. In addition, the data is created collaboratively by members of the Freebase community. The Freebase project has been officially stopped since 2016, but provided data can be accessed using API, RDF endpoint, and a full database dump.

Farsbase [8] is the first multi-source knowledge graph especially designed for the Persian language. Farsbase includes more than 500,000 entities with 7 million relations between them. Farsbase extracts its information from structured Wikipedia information such as infoboxes, web tables, as well as information extracted from the text extraction module. Farsbase also supports the ontology tree structure and uses the DBpedia knowledge graph to create its ontology.

CN-DBpedia [16] is a never-ending Chinese knowledge extraction system which is constantly updating knowledge by an end-to-end fact extraction model. Cn-DBpedia uses an update strategy by monitoring the changes of entities. Cn-DBpedia contains more than 10 million entities with 88 million relations between them. Accessing knowledge base is possible using API calls. Statistics show 164 million API calls which justifies success and importance of their system.

Zhishi.me [17] is a large scale Chinese Linking Open Data (LOD). Zhishi.me identifies structural features from three largest Chinese encyclopedias to extract data. This knowledge base has more than 5 million distinct entities. Zhishi.me allows users to run queries on data using SPARQL endpoints.

Recent developments in NLP systems in different languages, increases the need for knowledge graph construction. Ahmed et al. [18] introduce Arabic knowledge graph system. Like Persian, Arabic is a right-to-left language in which the sentence structure is different from English. These differences fail most NLP systems trained on English knowledge graphs.

Perez et al. [19] proposed a four stage knowledge graph construction from spanish textual resources. The experiments over the general and computer science domain show improvements in the identification of entities and relations. Spanish DBpedia was used to consider entities of resources.

Marchand et al. [20] presented a machine learning approach that extracts a knowledge graph from unstructured French documents. Cultural heritage in Quebec as a collection of French documents was used as a source of knowledge extraction in this project. Entity recognition in texts with domain-related types is also one of their enhancements.

3. FarsWikiKG Construction

FarsWikiKG stored data in RDF format. Persian Wikipedia is used as input and our goal is to convert these Wikipedia pages into RDF triples. Knowledge graph construction mainly consists of six modules as shown in Fig 1.

3-1. Extracting information

Wikipedia infoboxes have been used to build the knowledge graph in many previous knowledge graphs, such as YAGO and DBpedia. An infobox is a rectangular area in the upper left corner of some Wikipedia pages and consists of the main information of the corresponding page. The information in the infoboxes is in the form of pairs. For example, Fig 2 shows the Thomas Edison infobox where “occupation” is the key, and “Inventor” is the value. By extracting those pairs for the target entity, we achieve the desired triplets; e.g., (Thomas Edison, occupation, Inventor). In some relations like “children" each instance of the entities, each child in this example, will form a single triple in our knowledge graph linked to the main entity.

3-2. Cleaning and unifying information

The primary purpose of this section is to unify and organize the collected information to improve the efficiency and performance of the knowledge graph. In this part, two main tasks are done:

- Unification of equivalent relations: Two equivalent relations means two relations that have the same meaning; e.g., “birthplace” and “place of birth”. The existence of these relationships caused the mismatch problem in the knowledge graph information. A list of these equivalent relationships was prepared manually and the relations are mapped based on the list.

  Unification of complementary relations: Two complementary relations are two relations that can be achieved by inverting the concept of one to the other one; e.g., “father” and “child”; i.e., if we know “X” is the father of “Y”, then we can conclude that “Y” is the child of “X”. Therefore, it is not necessary to have both of these relationships together in the knowledge graph because it is possible to move from one relationship to another. Eliminating one side of these relationships can reduce the volume and improve the performance of the knowledge graph search.

3-3. Storing data in the database

In this step, the extracted triplets are stored in the database, neo4j. This database has been selected due to providing capabilities such as: receiving input information in triple form,

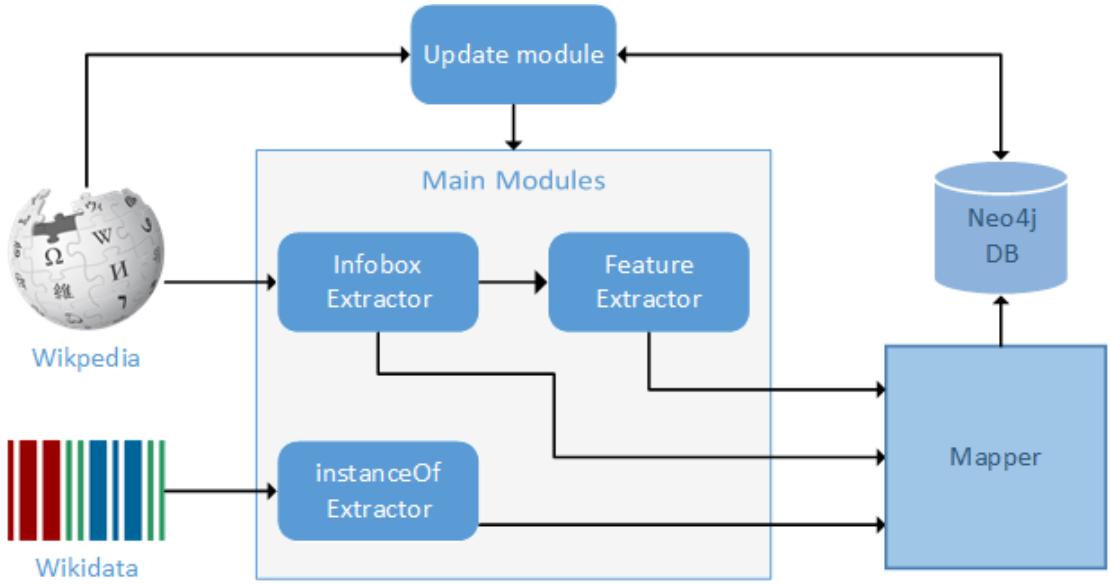

Fig. 1. FarsWikiKG construction modules

Fig. 2. Wikipedia infobox of "Thomas Edison"

the ability to execute queries, the ability to store metadata, having a graphical interface, and having rich libraries. An example of stored data shown in Fig 3.

3-4. Constructing ontology

FarsWikiKG uses Wikidata to build the ontology behind the knowledge graph. Wikipedia pages can be linked to Wikidata items and include data from Wikidata. Each Wikipedia page has "instance of" property in Wikidata, which contains detailed information and follows a hierarchy; e.g., Tehran is classified in Wikidata as "city of Iran", which itself is a subset of the "city/town". Therefore, the Wikidata "instance of" property uses a hierarchical structure which can be used to build a tree where each leaf is a node in the knowledge graph.

3-5. Maintaining metadata

In FarsWikiKG, entities include metadata. Some of these features are extracted in the fact extraction module; e.g., we

identify if an entity is associated with a specific page in Wikipedia or not.

The other metadata which is captured in this step is the type of entities which help the QA module to better search the knowledge graph and find the required information. Similar to the ontology, Wikidata is used to extract type features. Having ontology and the list of main types which are frequently used in QA systems, we manually select a number of nodes in this hierarchy, and for each entity, if the selected node is in the path to the root, we hold it as a label. The label is a more general form of the type we extracted before.

In addition to the metadata that are extracted directly from Wikidata, we also infer some types. This is mainly used for entities which do not have a specific page on Wikipedia; e.g., if an entity has no Wikipedia page, but contains specific relations which are dedicated to humans such as "date of birth", we can infer that its type is "human". A similar inference can be done to determine the gender of human entities, if the entity is identified as a father or mother of another person.

3-6. Updating the knowledge graph

Since Wikipedia pages are constantly changing, it is essential to have a separate module for updating. FarsWikiKG periodically updates itself and use new information to respond. For updating, we use an API which returns a list of recently

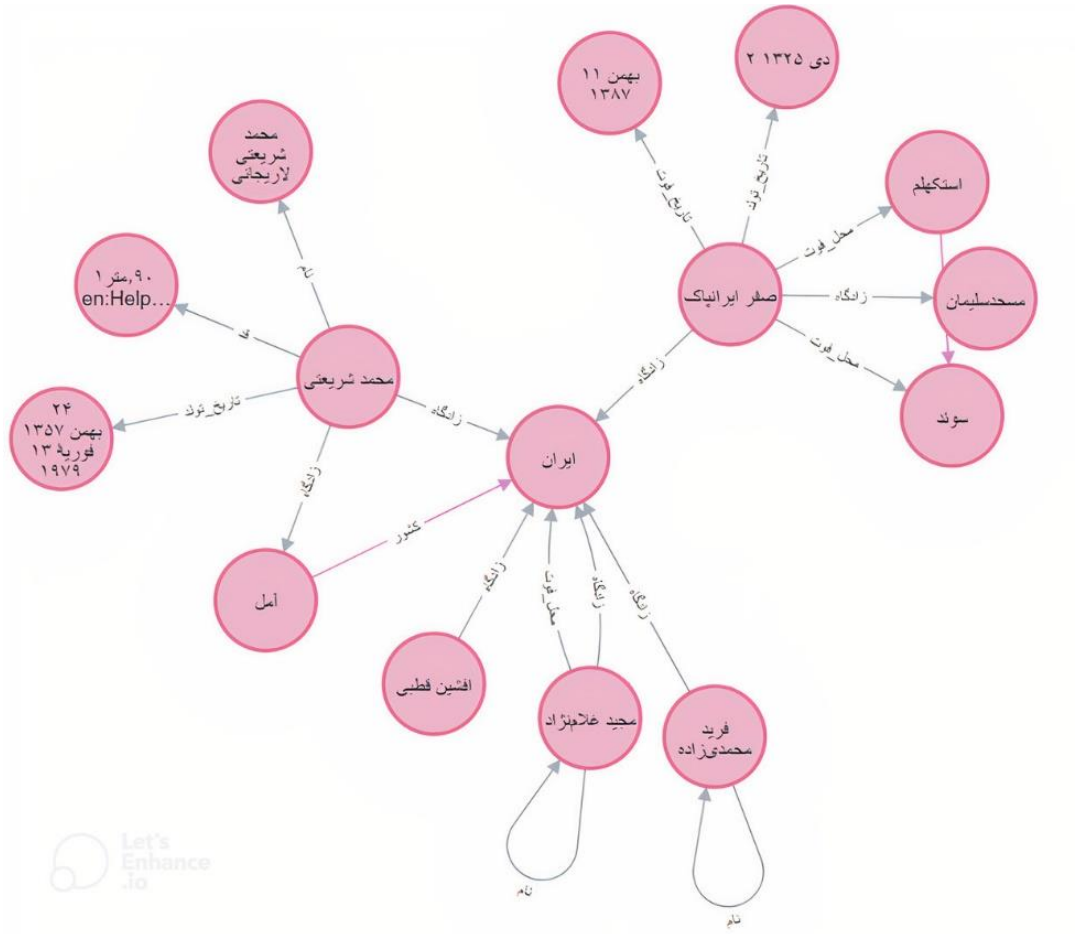

Fig. 3. Part of the stored knowledge graph

changed pages. Based on this list, the set of recently changed pages are selected and the update process is performed for the corresponding entities and relations.

4. Evaluation

In this paper, three different criteria have been used to evaluate the knowledge graph. Each of these criteria will be discussed below.

4-1. Size of the knowledge graph

The size of the knowledge graph are the main factors in evaluating the knowledge graph. This criterion examines the total volume of information in the knowledge graph. To evaluate this criterion, factors such as the number of entities, the number of relationships, number of entities with pages[2], number of entities with ontology, length of the largest ontology tree, and the number of different leaves of the ontology tree were measured. The results of this evaluation have been presented in **Error! Reference source not found.**Table I.

Comparing these results with Farsbase [8], the only available knowledge graph for Persian, we can see that FarsWikiKG has almost 4 times more entities compared to Farsbase. More entities count shows more information coverage of FarsWikiKG which we explain in more detail in the next experiment.

4-2. Information coverage of the knowledge graph

Knowledge graph coverage means that the knowledge graph can meet the information needs of users in various fields. Two different methods are used to evaluate knowledge graph coverage.

- Using the Wikipedia site: To evaluate this section, using the API of Wikipedia random section, 600 Wikipedia pages are randomly selected to check if their corresponding node is available in the knowledge graph. The result of this evaluation is reported in the first row of Table II.
- Using nominal entities: The purpose of this evaluation is to examine the existence of famous entities in the knowledge graph. These entities are the named entities that are searched frequently in search engines or QA systems. Most of these entities fall into three categories: person, location, and organization. To evaluate FarsWikiKG in terms of named entity coverage, the Peyma dataset [21] has been used. In the evaluation, 50 entities from each category are randomly selected, and then it is checked whether there is a node for the relevant entity. The results of this evaluation are reported in Table II.

[2] Entities that have a page in Wikipedia

TABLE I. THE STATISTICS OF FARSWIKIKG

#Relations	5770190
#Entities	2208123
#Entities with ontology	535295
#Entities with pages	531836
#Different leaves of the ontology tree	6655
Length of the largest ontology tree	15

TABLE II. COVERAGE OF FARSWIKIKG ON WIKIPEDIA PAGES AND NAMED

Type		#items	Coverage (%)
Wikipedia pages		600	89.83%
Named entities	Person	50	92%
	Location	50	92%
	Organization	50	94%

4-3. Accuracy and quality of the knowledge graph

In this section, the accuracy of the knowledge graph data is assessed more accurately, and attempts are made to assess this aspect more appropriately with several questions. Question answering systems as one the most usage of knowledge graphs, are great sources to evaluate quality of FarsWikiKG. In evaluation, in some relations like “children” each child will be linked to a single fact in our knowledge graph. In this section, three expert observers have asked to submit a set of suggested questions. Among these 100 total questions, FarsWikiKG can answer 96% of questions correctly which indicates the quality of this knowledge graph.

5. CONCLUSION AND FUTURE WORK

This paper presented FarsWikiKG, a Persian knowledge graph that uses Wikipedia infoboxes as the main source of extracting knowledge. FarsWikiKG combined Wikidata ontology tree with extra metadata helping QA systems to work properly.

Due to changes in Wikipedia and Wikidata, we update the knowledge graph regularly. Future work includes providing an entity linking and a QA system based on the features of our database.

REFERENCES

[1] P. Rajpurkar, J. Zhang, K. Lopyrev and P. Liang, "SQuad: 100,000+ questions for machine comprehension of text," Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 2383-2392.

[2] M. Joshi, E. Choi, D. S. Weld and L. Zettlemoyer, "TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension," Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, vol. 1, 2017, pp. 1601-1611.

[3] Z. Huang, S. Xu, M. Hu, X. Wang and J. Qiu, "Recent trends in deep learning based open-domain textual question answering systems," IEEE Access, vol. 8, pp. 94341-94356, 2020.

[4] S. Shin, X. Jin, J. Jung and K.-H. Lee, "Predicate constraints based question answering over knowledge graph," Information Processing and Management, vol. 56, pp. 445-462, 2019.

[5] X. Huang, J. Zhang, D. Li and P. Li, "Knowledge graph embedding based question answering," Proceedings of the 12th ACM International Conference on Web Search and Data Mining, no. Ccl, pp. 105-113, 2019.

[6] D. Lukovnikov, A. Fischer and J. Lehmann, "Pretrained Transformers for Simple Question Answering over Knowledge Graphs," In International Semantic Web Conference, Springer, Cham, 2019, pp. 470-486.

[7] Q. Guo, F. Zhuang, C. Qin, H. Zhu, X. Xie, H. Xiong and Q. He, "A Survey on Knowledge Graph-Based Recommender Systems," IEEE Transactions on Knowledge and Data Engineering, pp. 1-1, 2020.

[8] M. Asgari-Bidhendi, A. Hadian and B. Minaei-Bidgoli, "Farsbase: The persian knowledge graph," Semantic Web, vol. 10, p. 1169–1196, 2019.

[9] F. M. Suchanek, G. Kasneci and G. Weikum, "Yago: a core of semantic knowledge," in Proceedings of the 16th international conference on World Wide Web, 2007.

[10] J. Hoffart, F. M. Suchanek, K. Berberich and G. Weikum, "YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia," Artificial Intelligence, vol. 194, p. 28–61, 2013.

[11] F. Mahdisoltani, J. Biega and F. Suchanek, "Yago3: A knowledge base from multilingual wikipedias," in 7th biennial conference on innovative data systems research, 2014.

[12] T. Pellissier Tanon, G. Weikum and F. Suchanek, "Yago 4: A reason-able knowledge base," in European Semantic Web Conference, Springer, Cham, 2020, pp. 583-596.

[13] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer et al., "Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia," Semantic web, vol. 6, p. 167–195, 2015.

[14] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," Communications of the ACM, vol. 57, p. 78–85, 2014.

[15] 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, 2008, pp. 1247-1250.

[16] B. Xu, Y. Xu, J. Liang, C. Xie, B. Liang, W. Cui and Y. Xiao, "CN-DBpedia: A never-ending Chinese knowledge extraction system," in International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Springer, Cham, 2017, pp. 428-438.

[17] X. Niu, X. Sun, H. Wang, S. Rong, G. Qi and Y. Yu, "Zhishi. me-weaving chinese linking open data," in International Semantic Web Conference, Springer, Berlin, Heidelberg. 2011, pp. 205-220.

[18] I. A. Ahmed, F. N. AL-Aswadi, K. M. G. Noaman and W. Z. Alma'aitah, "Arabic Knowledge Graph Construction: A close look in the present and into the future," Journal of King Saud University - Computer and Information Sciences, 2022.

[19] A. G. García-Pérez, A. B. Ríos-Alvarado, T. Y. Guerrero-Meléndez, E. Tello-Leal and J. L. Martínez-Rodríguez, "An Approach for Knowledge Graph Construction from Spanish Texts," Research in Computing Science, vol. 149, p. 9–17, 2020.

[20] 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, Cham, 2020.

[21] M. S. Shahshahani, M. Mohseni, A. Shakery and H. Faili, "PAYMA: A Tagged Corpus of Persian Named Entities," Signal and Data Processing, vol. 16, pp. 91-110, 2019.

Saeedeh MomtazI is currently an associate professor at Amirkabir University of Technology (AUT), Iran. She completed her BSc and MSc education at Sharif University of Technology, Iran. She received a PhD degree in Artificial Intelligence from Saarland University, Germany. As part of her PhD, she was a visiting researcher at the Center of Language and Speech Processing at Johns Hopkins University, US. After finishing the PhD, she worked at the Hasso-Plattner Institute (HPI) at Potsdam University, Germany and the German Institute for International Educational Research (DIPF), Germany as a

postdoctoral researcher. Natural language processing is her main research focus.

Farhad Shirmardi received his BSc Computer Science at Amirkabir University of Technology (AUT), Iran in 2018. He received his MSc in Artificial Intelligence from Amirkabir University of Technology (AUT), Iran. His research interesets are Question Answering and Knowledge Graphs.

Mohammad Hadi Hosseisni He received his bachelor's degree in computer engineering from Amirkabir University of Technology (AUT), Iran in 2021. His research interests include machine learning, and algorithms, natural language processing.