IJICTR
International Journal of Information & Communication Technology Research

Volume 12- Number 4 – 2020 (60 -69)

A Graph-based Approach for Persian Entity Linking

Majid Asgari-Bidhendi*
Computer Engineering School
Iran University of Science and Technology
Tehran, Iran
majid.asgari@gmail.com

Farzane Fakhrian*
Computer Engineering School
Iran University of Science and Technology
Tehran, Iran
farzane.fakhrian@gmail.com

Behrouz Minaei-Bidgoli†
Computer Engineering School
Iran University of Science and Technology
Tehran, Iran
b_minaei@iust.ac.ir

Received: 10 September 2020 - Accepted: 12 November 2020

Abstract— **Most of the data on the web is in the form of natural language, but natural language is highly ambiguous, especially when it comes to the frequent occurrence of entities. The goal of entity linking is to find entity mentions and link them to their corresponding entities in an external knowledge base. Recently, FarsBase was introduced as the first Persian knowledge base with nearly 750,000 entities. This research suggested one of the first end-to-end unsupervised entity linking systems specifically for Persian, using context and graph-based features to rank candidate entities. To evaluate the proposed method, we used the first Persian entity-linking dataset created by crawling social media text from some popular Telegram channels. The ParsEL results show that the F-Score of the input data set is 87.1% and is comparable to any other entity-linking system that supports Persian.**

Keywords—Unsupervised Entity Linking; Entity Disambiguation; Persian Language; FarsBase; Knowledge Graph; Social Media Corpus

I. INTRODUCTION

Entity linking (EL) is the task of linking mentioned entities in the text with a knowledge base (KB). Entity linking is a developing area in natural language processing (NLP) and plays an essential role in text analysis, information extraction, questions answering, and recommendation systems [1]. It also enables users to know the prior knowledge of the entities in the text [2]. However, there are two types of ambiguity that make this task difficult. First, entities can have different names, even in a single document. For example, a person's name can appear in the text as a first, last, or nickname. EL needs to tie all of these names to a single entity in the knowledge base. Second, different entities can have the same name, but the entity linking system must be able to reference them to multiple entities in the knowledge base. Therefore, information about the entities is crucial in order to select the correct entities [2, 3, 4].

Fig. 1 presents a sample of disambiguation. For the word “apple”, the mention of Apple entity can refer to multiple entities; however, only one of them refers to the correct entity. In almost all cases, based on the information provided in the context, only one of the candidate entities can be correct.

The knowledge base is one of the basic components of the entity linking system. Generally, a knowledge base consists of a set of entities, information, semantic categories, and relationships between entities. The knowledge base used in the EL system must have some

* M. Asgari and F. Fakhrian contributed equally to this work as first authors.

† Corresponding author

[Downloaded from ijict.itrc.ac.ir on 2024-10-12]

features, such as public availability, Machine readability, persistent identifiers, and credibility [5].

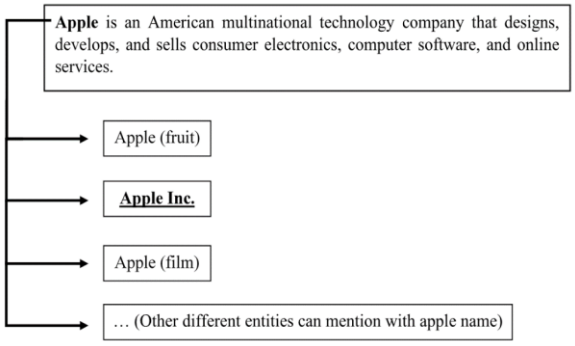

Fig. 1. Entity disambiguation for entity mention apple in a text. The correct entity is underlined.

There are currently several knowledge bases for EL systems, such as DBpedia [6], YAGO[1] [7], Freebase [8], and Probase [9]. This research uses FarsBase [10], which is the first multi-source knowledge base specifically designed for Persian, containing around 750,000 entities with 25 million relationships between them. FarsBase can provide various information, such as location, persons, and organization. FarsBase provides us with canonical data with a mapping system similar to DBpedia. ParsEL was originally developed by the FarsBase project and used some additional data not available on Wikipedia. For example, all villages are instances of the Village class in FarsBase, but the corresponding Wikipedia entry for each entity has different types of information boxes[2]. ParsEL uses class information in its heuristics. In addition to information about classes and related heuristics, this method can also work with Wikipedia as its knowledge base.

As mentioned earlier, Entity linking is a key part of many NLP application. Most of the previous studies were developed for situations where annotated training data is available, but this is not the case in many areas [11].

The biggest challenge when linking entities in Persian, besides the lack of sufficient resources for supervised learning, is the volume of content of each entity on Wikipedia and FarsBase. Because, the Persian articles are shorter than the English articles in different knowledge bases. In addition, many ancillary resources such as Wordnet, word embedding, BERT models, etc. are smaller in Persian or do not exist at all. And To the best of our knowledge, there were few entity linking system in Persian [12] before this study.

In this article, we proposed one of the first specially designed entity linkers for the Persian language and, to test our method, introduced the ParsEL-Social dataset, the first Persian entity linking dataset that can be used to link Persian entities in every entity linking system. Our systems achieve state-of-the-art performance compared to a few existing Persian entity linking systems.

The remainder of this article is organized as follows: Section 2 discusses the primary studies of entity linking; Section 3 presents the new EL dataset for the Persian language; Section 4 explains our knowledge base and Section 5 describes the proposed approach to entity linking in the Persian language. The results obtained with the baseline method are discussed in Section 6. The last section concludes this investigation and expresses our future work.

II. Related Works

In most cases, the entity-linking process has four subtasks that are consistent with most entity-linking systems. Fig. 2 shows this process and the sequence of these subtasks.

A. *Entity Recognition*

Most of the entity-linking studies [13,14,15] used existing entity recognition algorithms provided by other research and focused on the other three modules.

B. *Candidate Entity Generation*

Candidate entity generation is important task in EL process because a more accurate candidate entity generator can improve the whole linking process and EL system efficiency [16,17].

This module suggests a number of candidate entities for each entity mentioned in the text [3,18]. Most of the studies [2,19,20,21] used features such as redirect pages, disambiguation pages, and hyperlinks on Wikipedia or mean relation in YAGO to create a dictionary of names. For each entity mention, the name dictionary, map the entity mention to a set of candidate entities. It is also possible to get the set of candidates from this dictionary of names or from different surface form of entities in local documents [3].

Some linking systems [22,23,24] use web search engines and web information to find candidates. This method uses the top results from the selected search engine for the query with the entity mention as candidates.

Researchers [23,24,25] sometimes combine these methods and derive a list of candidates from the combination of results from different methods [3].

C. *Candidate Entity Ranking*

In most cases, the candidate entities are more than one. Therefore, the EL system has to classify the candidate entities in order to find the appropriate entity in the knowledge base [5]. The EL system can use two types of features to rank candidate entities, namely, context-independent features and context-dependent features [3]. In the literature, the term "phase entity disambiguation" [22,26,27] has the same meaning as the candidate entity ranking. In addition, supervised and unsupervised methods can be used to achieve the results. Supervised methods depend on the annotated training dataset and its data annotation must be done manually.

[1] YAGO (Yet Another Great Ontology)

[2] Wikipedia Info Box

[Downloaded from ijict.itrc.ac.ir on 2024-10-12]

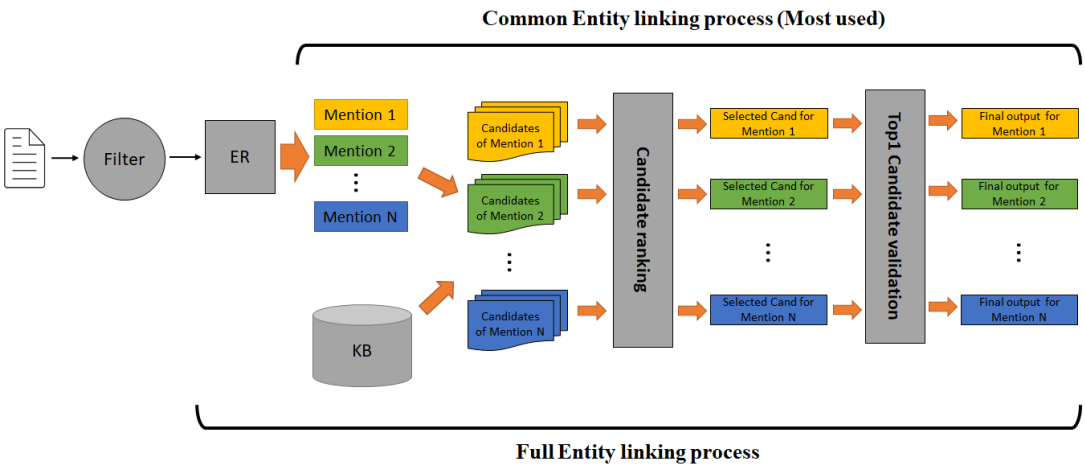

Fig. 2. Entity Linking process

In some studies [15,26] the researchers publish their manually annotated dataset for EL. These datasets are excellent benchmarks for the entity linking task. However, in the social media field, creating such a dataset is very difficult, costly, time-consuming, and moreover, most of the studies in EL focus on the English language. Because of such a shortcoming, we decided to work in unsupervised methods.

Some researchers [18,23,26,28,29] used methods based on the vector space model (VSM) [30] to unsupervised candidates ranking. In this process, the first step is to compute the similarity between the vector representations of the entity mention and the candidate entity. The system links the candidate entity with the highest similarity to the entity mention. Their methods differ in the calculation of the vector similarity and the vector representation [3]. In addition, working on social media makes work difficult due to colloquial text and spelling issues.

Cucerzan used entity references mentioned in context and candidate entity articles to create vectors. To do this, the system selects a candidate that maximizes vector similarity and has the same category as an entity mention. This system achieved an accuracy of 91.4% in a news dataset [26].

Chen et al., Constructed the entity mention and candidate entity vectors based on the Bag of Words model using the context of their article to capture information about the co-occurrence of words and calculate the similarity between them using TF-IDF similarity. They stated an accuracy of 71.2% in the TAC-KBP2010 dataset [28].

Han and Zhao used two types of similarity measures: the similarity based on the semantic knowledge base of Wikipedia together with the similarity based on Bag of Words method. To generate vectors in the semantic knowledge base of Wikipedia similarity, the method detects Wikipedia concepts in candidate entities and the context of the mentioned entity, and then computes the vector of Entity mention and candidate entities similarities using a weighted average of semantic relations between the concepts of Wikipedia articles and the context of the mentioned entity. Thereafter, these two types of similarity are merged and the final similarity vector of the candidate entities is reported and ultimately the entity that maximizes this merged similarity is selected. Their system achieves an accuracy of 76.7 in the data set TAC-KBP2009 [23].

Nozza et al., used the Word Embeddings representation instead of the Bag of Words representation in the Micropost 2015 and 2016 datasets which contain different Tweets. In their method, first, they narrow down the candidates using the similarity between the entity mention and each candidate entity, and second, they use skip-gram as word embedding to score the remaining candidates. They reported 53% precision in these two datasets. Their results are very low compared to other EL systems because they have worked on social media texts and, as already explained, user-generated texts represent a greater challenge than others [16].

Xu et al., applied a linkage approach to medical texts, taking advantage of name similarity, entity popularity, category consistency, context similarity, and semantic correlation between entity mention and candidate entities, and rank these candidates by combining these features. They call their ranking measure the confidence score. On average, this confidence score in medical dataset is 82% accurate [29].

Zhang et al., proposed an unsupervised bilingual entity linker inspired by the works of Han and Sun [31] and Yamada, Shindo, Takeda, and Takefuji [27]. As mentioned earlier, they used a pre-made dictionary for candidate generation, and after that they used probabilistic generative methods to disambiguate entities. Their system achieves an accuracy of 91.2% in the CoNLL dataset [32].

Xie et al., used structured relationships between entities in their local knowledge base and background data from other knowledge bases and improved the weighted method of Word2Vec and PageRank for their similarity assessment. They named their method Graph Ranking Collective Chinese Entity Linking (GRCCEL) and reported an accuracy of 88.12 in the Sogou-NED corpus in Chinese [33].

[Downloaded from ijict.itrc.ac.ir on 2024-10-12]

TABLE I. RESULTS OF MAIN RESEARCHES THAT USED UNSUPERVISED LEARNING FOR RANKING.

Research	Methods for unsupervised candidate ranking	Precision	Dataset
Nozza et al. [16]	Used Similarity score and skip-gram word embedding	53%	Micropost 2015 and 2016 (tweets)
Chen et al. [28]	Construct vectors based on Bag of Words and TF-IDF similarity score	71.2%	TAC-KBP2010
Han and Zhao [23]	Merge two similarity method based on Wikipedia knowledge base and Bag of Words method	76.7%	TAC-KBP2009
Xu et al. [29]	Use name similarity, entity popularity, category consistency, context similarity, and semantic correlation features to rank candidates	82%	Online Chinese Medical Text
Xie et al. [33]	Improve the weighted method of Word2Vec and PageRank for their similarity assessment	88.12%	Sogou-NED
Zhang et al. [32]	Use probabilistic generative method	91.2%	CoNLL
Cucerzan [26]	Create vectors used entity references mentioned in context and calculate vector similarity	91.4%	News dataset
Pan et al. [35]	Used Abstract Meaning Representation (AMR) and graph methods	92.12%	News and discussion forum posts dataset

Pan et al., used Abstract Meaning Representation (AMR) [34] to select high quality entity sets for their similarity measure. They indicated that their representation using AMR could capture some contextual properties that are very critical and useful for entity disambiguation without using training data. To compare the entities context, they next used an unsupervised graph to get final results and reported 92.12% accuracy for an annotated dataset from news and discussion forum posts [35].

Table 1 summarizes the final results of some of the previous studies in unsupervised entity linking. Researchers have done less research on entity linking in the Persian language than it has in English, because we face major challenges in linking Persian entities like lack of sufficient resources, proper dataset and short articles in existing knowledge bases. And as far as we know, the proposed solution is one of the earliest Persian entity linkers.

Babelfy[1] [36] is one of the entity linking systems that works on Persian and works based on BabelNet 3.0[2] knowledge base. Babelfy uses a three-step method for entity disambiguation: first, a semantic signature is created for each node (concept or named entity), which is actually a set of nodes that are connected to that node in the main BabelNet graph. Similar to other systems, Babelfy gets all possible candidates, which are the nodes inside the BabelNet, by entering the text for each entity mention. And then use semantic signatures, a sub-graph of the BabelNet is generated, and the disambiguation process is done using a centrality measure. We use Babelfy as our baseline method.

D. Unlinkable Mention Prediction

In cases where entity mentions do not have relevant entities in the knowledge base, unlinkable entity mentions can be separated from other entities and marked as NIL. Researchers have suggested several ways, i.e. ignoring mentions of unlinkable entities, [2,26,31] NIL threshold, [3,27] and supervised machine learning techniques like binary classification methods [3,5,37,38] to separate unlinkable mentions and return correct candidate for entity mention. Some studies called this subtask "Top 1 Candidate Validation".

III. DATASET

In this research, we use the ParsEL-Social corpus, the first Persian entity linking dataset. To build the corpus, we selected 10 categories (sport, economics, gaming, general news, IT news, travel, art, academic, entertainment, and health.) and one telegram channel for each category. Crawler crawled the posts of each channel one by one and added the entire post to the corpus. The adding process continued until the sum of the number of words in the category exceeds X. This X was the same for all the channels. After the crawling phase, we finally realized that some posts were advertising in nature and were repetitive. Duplicate posts were removed from the corpus, and only one copy was retained. Therefore, all of the posts in the corpus are not incomplete.

In the annotation process, ParsEL software linked the entities on all of the posts in the corpus and generated the candidates for each mention. This initial process was executed to facilitate the work of experts in the system. Experts edited the gold links in a user interface. The system allows the experts to set any link even outside of the candidate list for a token. The work of each expert has been verified by at least another expert. We have not set any limitations on the classes, and the corpus has a link of any entities even from the Thing class.

Table 2 summarizes the statistics of the ParsEL-Social dataset, such as the number of posts, words and entities, the average number of words and entities in each post, and the average number of candidates for each entity.

Moreover, table 3 shows the statistics of the ParsEL Social dataset for each category. Should be noted, the

[1] http://babelfy.org

[2] https://babelnet.org

[Downloaded from ijict.itrc.ac.ir on 2024-10-12]

numbers of documents in the sport and academic categories are higher than in other categories because the posts are shorter in the datasets. The number of sentences is fewer in the game, travel, and health categories because longer sentences are used in the posts. The corpus distributes an equal number of words for all the categories. Texts in the sport, general news, and academic categories have a higher number of entities, but the differences are not remarkable as the entities are not restricted to named entities. Finally, the number of candidates is higher in the sport, general news, and academic categories; therefore, these types of texts have more ambiguous words.

TABLE II. PARSEL-SOCIAL DATASET PROPERTIES

Dataset	Count
Documents	4,263
Sentences	6,160
Words	67,595
Entities	19,831
Candidates	145,148
Words per article	15.9
Entities per article	4.7
Candidates per Entity mentions	7.3

IV. KNOWLEDGE BASE

FarsBase is a knowledge base with several input sources that are made exclusively for the Persian language. Like DBpedia and many other knowledge bases, the mainstay of the Persian knowledge base is Wikipedia. In addition to Wikipedia, FarsBase extracts knowledge from web tables and raw texts. Along with FarsBase, a search engine is provided that can be used to search for knowledge using natural language queries. One of the most important forms of search in this system is the search for entities.

One of the most important parts of raw text extractor in FarsBase is the module of entity linking. According to the process of knowledge extraction, the knowledge in the text must be produced in RDF triples, therefore the entities within the text must be identified. Also, considering that the subject (in RDF format) must be linked to one of the entities of the knowledge base, the entity linking operation plays a very important role in knowledge extraction. ParsEL is also used in response to search engine queries. If the query phrase is exactly equivalent to an entity, that entity is returned as the result. And if a predicate is requested from an entity (for example, Ali Daei's height), the entity is first identified and then the object that is related to the entity through the predicate is returned. Also, in the process of extracting knowledge from web tables, the first step is to recognize and link the entities within the cells [10].

V. PROPOSED ENTITY LINKING METHOD (METHODOLOGY)

Like other entity linking methods, the proposed method focuses on the candidate generation, the ranking, and unlinkable mention predictions.

Initially, the proposed method uses FarsBase for the candidate entity generation. For each entity in FarsBase, a predicate named “variantLabel” obtains its values from Wikipedia redirect pages and has different versions of the name of entities. For every word, the algorithm extracts all possible entities based on its variantLabel in the FarsBase. By using this method, we can generate a candidate set for each entity mention in the candidate ranking of the next step.

In the candidate ranking phase, the goal is to link each entity mention to only one knowledge base entity from the candidate set. We utilize both of the context-dependent and context-independent features in the ranking step. Context-dependent features rely on the context where entity mention appears, but context-independent features are independent of context and rely on entity mention and candidate entities [3].

TABLE III. PARSEL-SOCIAL DATASET STATISTICS PER CATEGORY

	sport	economy	game	IT-News	General-News	travel	art	academic	fun	health
Number Of Documents	469	292	734	383	362	331	389	457	379	467
Number Of Sentence	680	595	866	515	583	539	572	737	531	542
Number Of Words	7564	7495	5986	7282	8729	6272	6233	7136	5967	4931
Number Of Entities	2795	2212	1206	1865	2857	1669	1728	2715	1405	1379
Number Of Candidates	21920	15914	10263	14891	20210	11603	10769	16689	13365	9524

[Downloaded from ijict.itrc.ac.ir on 2024-10-12]

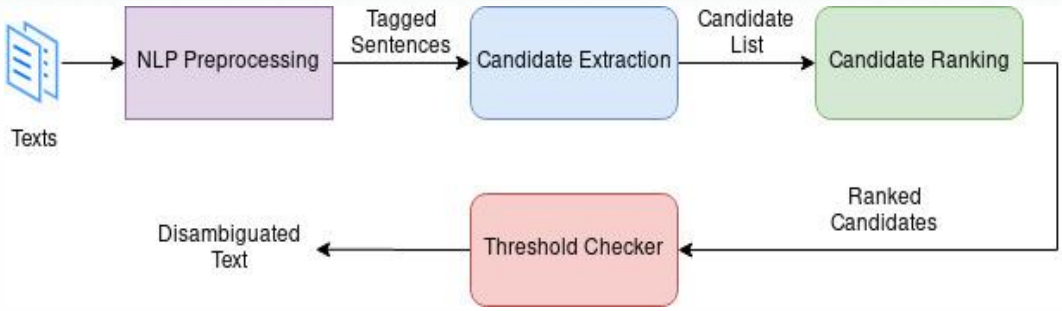

Fig. 3. Overview of ParsEL 1.0

First of all, we use these following heuristics to remove some of the inappropriate candidates:

a) Type Checking: The system checks the types of entities and eliminates candidates whose type is not the same as the entity mention of the candidate set.

b) POS Tags: Following the type checking, a built-in POS tagger from our knowledge base is used to tag sentences surrounding the entity mentions and eliminate the entities that have POS tag different from the entity mentions. We have used JHazm software for POS tagging [1] . In particular, ParsEL filters the preposition and do not link them to any FarsBase entities. In this heuristic, the role of the word must match with the class of the entity. For example, in Persian, "به" can be a noun (quince fruit) or a preposition (to). Also, "می‌رود", can be a noun (Meyrood village) or a verb (is going). Therefore, the POS tag can help ParsEL to filter some of the incorrect candidate entities.

c) The popularity of entities: Entities with the same mention have different popularity [3]. Take Tehran as an example; Tehran (city) is much more used than Tehran University. Therefore, in these cases, rare entities are ignored using a manually created list. FarsBase was developed to help users of Persian search engines and had restricted access to anonymous log queries of the users. In the current version, we do not have access to search logs anymore, and we assumed that longer articles are more popular (we can also use Wikipedia APIs[2] in the future as a better metric for popularity detection).

d) Class-specific Filters: Some entities have a very generic name that may cause a high level of ambiguity. For instance, "چهل سالگی" ("At the age of 40") is an Iranian movie while it can be as a part of a general sentence, e.g., "Vahid died at the age of 40". Such names are widespread in artworks (e.g., movies or books) and a limited number of the other specialized classes. To improve the disambiguation process, we look for more evidence in the context using a reference list if the candidate entity belongs to individual classes. Considering the above example, "At the age of 40", the surrounding context containing phrases such as channel, cinema, ticket, and a movie is required. Otherwise, the algorithm multiplies the real rate of the candidate by a predefined constant number between 0 and 1 based on each case. Currently, these filters are used only for Work class of FarsBase and all of its subclasses e.g. Movie, Series, Music Work.

After removing some of the incorrect candidates, the system scores the remaining candidates. The scoring method employs context-dependent features and follows the four following steps:

- **Context Score:** The first step is to compute the cosine similarity between the words of the context of the entity mention and the textual context of the corresponding Wikipedia article of candidate entities. This step ignores the stop words in the Persian language.
- **Graph Score:** In the next step, candidates are scored based on the number of hyperlinks between all candidate entities in their corresponding Wikipedia articles.
- To rank the candidates, we merge the context score and graph score. Each factor (graph and context) divides a score between zero and one among the candidates. The final score is obtained from the weighted sum of the factors.
- Finally, the system links the candidate entity with the highest score to the entity mention. Other entities will be added to the entity mention's "ambiguity-list" to persist the rejected candidates for possible future applications such as error checking.

After candidate generation and ranking, the NIL threshold method [3,27,39] is used for unlinkable mention prediction. In this method, if the score of the top-ranked candidate entity is lower than the pre-defined threshold, the entity mention is tagged as NIL, and the system adds all of the candidate entities to the ambiguity-list.

[1] an improved version of https://github.com/mojtaba-khallash/JHazm

[2] https://en.wikipedia.org/wiki/Wikipedia:Pageview_statistics

[Downloaded from ijict.itrc.ac.ir on 2024-10-12]

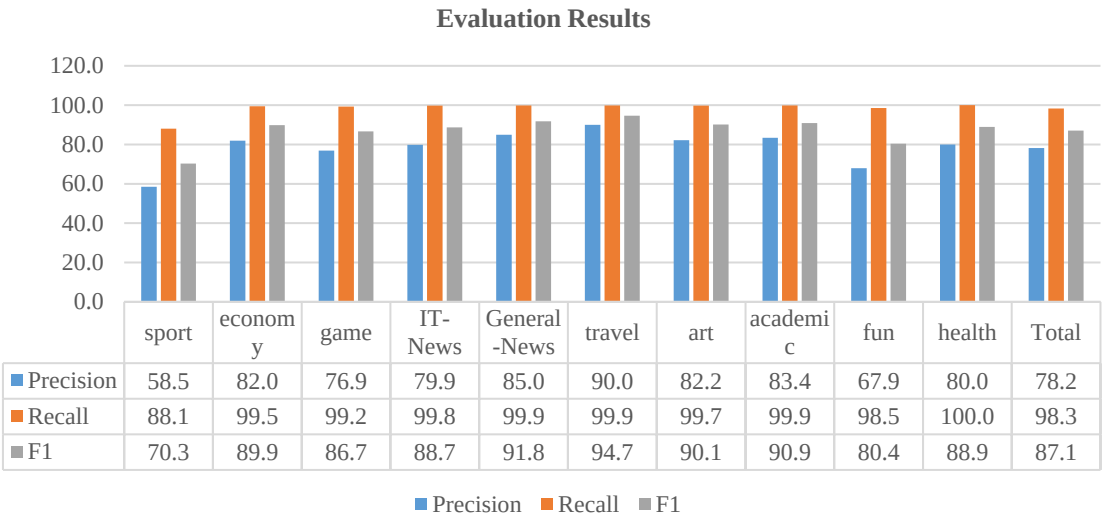

Fig. 4. Entity linking results using the proposed method on ParsEL-Social dataset.

TABLE IV. COMPARING PARSEL1.1 WITH PARSEL1.0 AND THE BASELINE ALGORITHM.

Category	Baseline P	Baseline R	Baseline F1	ParsEL 1.0 P	ParsEL 1.0 R	ParsEL 1.0 F1	ParsEL 1.1 P	ParsEL1.1 R	ParsEL1.1 F1
Sports	0.5111	0.4720	0.4908	0.5647	0.9282	0.7022	0.5855	0.8805	0.7033
Economy	0.4855	0.5676	0.5234	0.8174	0.9961	0.8979	0.8197	0.9946	0.8987
Game	0.3790	0.5430	0.4464	0.7669	0.9934	0.8656	0.7691	0.9925	0.8666
IT News	0.4061	0.4375	0.4212	0.7974	0.9994	0.8870	0.7985	0.9976	0.8870
General News	0.4638	0.5080	0.4849	0.8476	1.0000	0.9175	0.8495	0.9985	0.9180
Travel	0.4572	0.2297	0.3058	0.8946	1.0000	0.9444	0.8997	0.9987	0.9466
Art	0.4576	0.2746	0.3433	0.8193	0.9987	0.9002	0.8223	0.9975	0.9014
Academic	0.5757	0.5296	0.5517	0.8252	1.0000	0.9042	0.8339	0.9988	0.9090
Fun	0.4279	0.4531	0.4402	0.6707	1.0000	0.8029	0.6793	0.9853	0.8042
Health	0.4830	0.4818	0.4824	0.7987	1.0000	0.8881	0.8003	1.0000	0.8890
Total	0.4716	0.4546	0.4630	0.7744	**0.9911**	0.8694	**0.7817**	0.9831	**0.8710**

Suppose the input sentence " شیر به دنبال آهوی کوهی میدوید." (The lion runs after the mountain gazelle). An NLP preprocessing phase (including sentence boundary detection, work tokenization, and part of speech tagging) is applied to the input sentence. Candidate extractor module receives the processed sentences as its input and detects 20 different candidates for "شیر" which including "Lion (animal)", "Leo (constellation)", "Milk" and "Faucet". It also detects three candidates for "به" including Quince (fruit), one candidate for "آهوی کوهی" and four candidates for "کوهی". The system removes all candidates for "به" because it is a proposition based on the POS tags of the sentence. The system ranks all of 20 candidates of "شیر" and assigns a confidence value to each candidate based on the heuristics we mentioned above. In the next step, the system ranks them and selects the candidate with the most confidence. If the confidence of all entities is lower than 0.001, system removes all of them, and links the word to NIL. This operation is repeated for "آهوی" and "کوهی". Note that the system prefers longer entities to shorter ones and assigns a greater base factor to "آهوی کوهی" than other candidates of the word "کوهی". Base factor is multiplied with the confidence of each entity in the ranking phase. Fig. 3 depicts the workflow and the proposed method in pseudo-code form is shown in the appendix.

For improving results of candidates' disambiguation, we used named entity recognitions (NER) types. In this feature, a candidate's final grade increases if it is recognized as a person, location, or etc. in our NER system and increases the likelihood that the candidate will be elected. Moreover, our method suggests new candidate entities based on NER types. If an entity mention identity as a NER type (other than the MISC type) and its candidate list was empty, our system will suggest a new entity to the knowledge base. This method suggests 547 new entities to FarsBase. Our entity linker is available to the public[1]. And it can be used to link the entities in an input text.

[1] http://farsbase.net/parsel

[Downloaded from ijict.itrc.ac.ir on 2024-10-12]

TABLE V. NAMED ENTITIES NOT LINKED IN OUR SYSTEM

NER types	Total Count	ParsEL Not Tagged	Gold Not Tagged	ParsEL Percent	Gold Percent
NE	1197	107	151	8.939	12.615
LOC	1914	1279	1312	66.823	68.548
MISC	1860	411	446	22.097	23.978
ORG	1466	316	409	21.555	27.899

VI. RESULTS AND EVALUATION

We evaluate the proposed unsupervised method (ParsEL 1.1[1]) on the ParsEL-Social dataset, and the results are reported in Fig. 4 for each category using precision, recall, and F1 measures. The proposed method is comparable with the state-of-the-art unsupervised methods on TAC-KBP datasets, and the results are acceptable for the first Persian entity linker. Generally, EL in Persian is easier than English. Because English articles are 10 times more than Persian articles in FarsBase or Wikipedia. So, we have fewer candidates in-average for our entities. For example, in AIDA CoNLL-YAGO Dataset [15], the average count of candidates for each entity is 70. In Persian, some words don't even have a separate page in Wikipedia. So we have less ambiguity in Persian and that's the reason which made ParsEL Recall so high. Table 4 compares ParsEL1.1 results with ParsEL1.0 and Babelfy entity linking (our baseline method). And Table 5 shows the percentage of entities with a NER type that our entity linking system was unable to link.

ParsEL1.1 results compared to ParsEL1.0 is improved by using NER types in entity linking process. And suggest useful entities to KB.

In ParsEL and Babelfy comparison, both use graph-based methods to resolve the ambiguities. Unlike ParsEL, Babelfy does not use contextual information directly, i.e. the similarity distance between the words of the text and the description text of each node is not calculated and only graph-based methods are used. Although BabelNet includes WordNet synsets, the contextual information has been used indirectly, but this information cannot be used for Persian language sentences because of the lack of a public version of FarsNet (Persian WordNet). The second difference is that clustering and random walk mechanism are used to form the sub-graph, therefore the graph is constructed probabilistically. In ParsEL we have created the graph based on internal links of Wikipedia.

In the first step, we run Babelfy on our dataset by public APIs of Babelfy. Babelfy returns all of the BabelNet synsets for each token in the text. Each synset is linked to some sources such as Wordnet or Wikipedia articles in different languages. Synset sources are available on the page of the synset or public BabelNet APIs. Each Wikipedia article in the Persian language corresponds to a FarsBase entity. Since BabelNet merges multiple sources to construct its synsets and, on the other hand, FarsBase is based on Persian Wikipedia, we only get Persian Wikipedia sources for each synset and convert it to FarsBase links. Therefore, each BabelNet synset can be linked to its corresponding entity in the FarsBase knowledge graph. As it was discussed earlier, BabelNet synsets are not extracted only from Persian Wikipedia, thus, comparing the reported recall rate with the ParsEL is not wholly impartial, and it is normal for baseline recall to be lower. Anyway, Babelfy is the only system that worked on Persian, and the experiments can be executable via API or code. To choose the baseline, we considered two criteria: First, the code related to the method is available on the Internet, or all sections can be implemented from the study. Second, there should be an API for that baseline system (for example, Babelfy and Google Natural Language services). There were several options, but only one of them (Babelfy) supports the Persian language. There are significant differences between BabelNet and FarsBase. In short, FarsBase has been developed specifically for the Persian language and is a multi-source knowledge base and its entities are extracted only from Wikipedia articles, but BabelNet is multilingual and uses both Wikipedia and WordNet and other sources simultaneously. Babelfy is both an entity linker and a WSD system.

VII. CONCLUSIONS AND FUTURE WORK

In this paper, we presented ParsEL, an entity linker for the Persian language, which uses the FarsBase knowledge graph as its dataset. The results show that the precision of ParsEL is comparable with the entity linkers in other languages. Using multiple heuristics enables ParsEL to compete with state-of-the-art unsupervised methods for entity linking even in other languages.

In future work, we plan to annotate a larger dataset for supervised approaches. Deep learning has improved entity linking results in recent years, which can be the correct choice for the next versions of ParsEL. Besides, extracted links from a piece of text must have reasonable relationships. A post-processing phase can investigate these relationships and improve the overall results. Using entity or word embedding also can improve the proposed method for entity linking in the Persian language. BERT-based models were able to perform better than their predecessors in various natural language processing tasks, including named entity recognition. Unlike context-independent models such as word2vec, BERT is a context-dependent representation model. In most of the word-embedding methods which are devised before BERT, each word has a fixed vector, but in the BERT, word vectors are different in each context.

Multilingual-BERT (M-BERT) released by [40] is available in 104 languages including Persian. M-BERT has been trained with the largest Wikipedia. Research has been done exclusively to train the BERT model for the Persian language. Including ParsBERT [41] and SINA-BERT [42]. A model has also been developed in the data mining laboratory of the faculty of computer

[1] ParsEL is the entity linker Raw-Text Extractor Module of the FarsBase project. FarsBase is an open-source system and is available in https://github.com/IUST-DMLab/farsbase-kg.

[Downloaded from ijict.itrc.ac.ir on 2024-10-12]

engineering of the Iran University of Science and Technology (IUST), which is larger than the previously mentioned models.

We can target more challenging baselines like the models that have performed well on other languages such as English. Comparison of our method with other methods in English is possible in two ways: 1- Having a parallel Persian-English EL corpus and the evaluation of the methods on this corpus. Of course, this comparison is not fair because of the huge difference in the number of candidates for these two languages. 2- Our method should be ported and evaluated in the other language. We are working this approach currently (we named it ULIED - Unsupervised Language-Independent Entity Disambiguation). For the comparison in this case, we have used Wikipedia as the knowledge base, and we have eliminated the heuristics which require non-Wikipedia resources.

REFERENCES

[1] W. Yan and K. Khurad, "Entity Linking with people entity on Wikipedia," *arXiv Prepr. arXiv1705.01042*, 2017.

[2] X. Han, L. Sun, and J. Zhao, "Collective entity linking in web text: a graph-based method," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 2011, pp. 765–774.

[3] W. Shen, J. Wang, and J. Han, "Entity linking with a knowledge base: Issues, techniques, and solutions," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 2, pp. 443–460, 2014.

[4] O.-E. Ganea, M. Ganea, A. Lucchi, C. Eickhoff, and T. Hofmann, "Probabilistic bag-of-hyperlinks model for entity linking," in *Proceedings of the 25th International Conference on World Wide Web*, 2016, pp. 927–938.

[5] P. Taufer, R.M. Straka, "Named Entity Recognition and Linking," 2017.

[6] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *The semantic web*, Springer, 2007, pp. 722–735.

[7] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: A Core of Semantic Knowledge Unifying WordNet and Wikipedia," in *Proceedings of the 16th international conference on World Wide Web - WWW '07*, 2007, p. 697, doi: 10.1145/1242572.1242667.

[8] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," *SIGMOD 08 Proc. 2008 ACM SIGMOD Int. Conf. Manag. data*, pp. 1247–1250, 2008, doi: 10.1145/1376616.1376746.

[9] W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: A probabilistic taxonomy for text understanding," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, 2012, pp. 481–492.

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

[11] A. Arora, A. García-Durán, and R. West, "Low-rank Subspaces for Unsupervised Entity Linking," CoRR, vol. abs/2104.08737 ,2021.

[12] M. Asgari-Bidhendi, B. Janfada, A. Havangi, S.-A. Hossayni, and B. Minaei-Bidgoli, "An Unsupervised Language-Independent Entity Disambiguation Method and its Evaluation on the English and Persian Languages," CoRR, vol. abs/2102.00395, 2021.

[13] M. Pershina, Y. He, and R. Grishman, "Personalized page rank for named entity disambiguation," in Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 238–243.

[14] C. Ran, W. Shen, and J. Wang, "An Attention Factor Graph Model for Tweet Entity Linking," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 1135–1144.

[15] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum, "Robust disambiguation of named entities in text," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 782–792.

[16] D. Nozza, C. Sas, E. Fersini, and E. Messina, "Word Embeddings for Unsupervised Named Entity Linking," in Knowledge Science, Engineering and Management, 2019, pp. 115–132.

[17] B. Hachey, W. Radford, J. Nothman, M. Honnibal, and J. R. Curran, "Evaluating Entity Linking with Wikipedia," Artif. Intell., vol. 194, pp. 130–150, Jan. 2013.

[18] G. Wu, Y. He, and X. Hu, "Entity linking: an issue to extract corresponding entity with knowledge base," *IEEE Access*, vol. 6, pp. 6220–6231, 2018.

[19] S. Guo, M.-W. Chang, and E. Kiciman, "To link or not to link? a study on end-to-end tweet entity linking," in Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013, pp. 1020–1030.

[20] A. Gattani, D. S. Lamba, N. Garera, M. Tiwari, X. Chai, S. Das, S. Subramaniam, A. Rajaraman, V. Harinarayan, and A. Doan, "Entity extraction, linking, classification, and tagging for social media: a wikipedia-based approach," *Proc. VLDB Endow.*, vol. 6, no. 11, pp. 1126–1137, 2013.

[21] W.-H. Chong, E.-P. Lim, and W. Cohen, "Collective entity linking in tweets over space and time," in *European Conference on Information Retrieval*, 2017, pp. 82–94.

[22] M. Dredze, P. McNamee, D. Rao, A. Gerber, and T. Finin, "Entity disambiguation for knowledge base population," in Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), 2010, pp. 277–285.

[23] X. Han and J. Zhao, "NLPR_KBP in TAC 2009 KBP Track: A Two-Stage Method to Entity Linking.," 2009.

[24] S. Monahan, J. Lehmann, T. Nyberg, J. Plymale, and A. Jung, "Cross-lingual cross-document coreference with entity linking," in TAC 2011 Workshop, 2011.

[25] J. Lehmann, S. Monahan, L. Nezda, A. Jung, and Y. Shi, "Lcc approaches to knowledge base population at tac 2010," in TAC 2010 Workshop, 2010.

[26] S. Cucerzan, "Large-scale named entity disambiguation based on Wikipedia data," in Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), 2007, pp. 708–716.

[27] I. Yamada, H. Shindo, H. Takeda, and Y. Takefuji, "Joint learning of the embedding of words and entities for named entity disambiguation," *arXiv Prepr. arXiv1601.01343*, 2016.

[28] Z. Chen, S. Tamang, A. Lee, X. Li, W.-P. Lin, M. G. Snover, J. Artiles, M. Passantino, and H. Ji, "CUNY-BLENDER TAC-KBP2010 Entity Linking and Slot Filling System Description.," 2010.

[29] J. Xu, L. Gan, M. Cheng, and Q. Wu, "Unsupervised medical entity recognition and linking in Chinese online medical text," J. Healthc. Eng., vol. 2018, 2018.

[30] G. Salton, A. Wong, and C. S. Yang, "Avector space model for automatic indexing," Commun. ACM, vol. 18, no. 11, 1975.

[31] X. Han and L. Sun, "A generative entity-mention model for linking entities with knowledge base," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 945–954.

[32] J. Zhang, Y. Cao, L. Hou, J. Li, and H.-T. Zheng, "XLink: An unsupervised bilingual entity linking system," in Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, Springer, 2017, pp. 172–183.

[33] T. Xie, B. Wu, B. Jia, and B. Wang, "Graph-ranking collective Chinese entity linking algorithm," Front. Comput. Sci., vol. 14, pp. 291–303, 2019.

[34] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N.

[Downloaded from ijict.itrc.ac.ir on 2024-10-12]

Schneider, "Abstract meaning representation for sembanking," in Proceedings of the 7th linguistic annotation workshop and interoperability with discourse, 2013, pp. 178–186.

[35] X. Pan, T. Cassidy, U. Hermjakob, H. Ji, and K. Knight, "Unsupervised entity linking with abstract meaning representation," in Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics: Human language technologies, 2015, pp. 1130–1139.

[36] A. Moro, A. Raganato, and R. Navigli, "Entity Linking meets Word Sense Disambiguation: a Unified Approach," in Transactions of the Association for Computational Linguistics., vol. 2, 2014, pp. 231–244.

[37] W. Zhang, C. L. Tan, Y. C. Sim, and J. Su, "NUS-I2R: Learning a Combined System for Entity Linking.," 2010.

[38] W. Zhang, Y.-C. Sim, J. Su, and C.-L. Tan, "Entity linking with effective acronym expansion, instance selection and topic modeling," 2011.

[39] W. Shen, J. Wang, P. Luo, and M. Wang, "Linden: linking named entities with knowledge base via semantic knowledge," in Proceedings of the 21st international conference on World Wide Web, 2012, pp. 449–458.

[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," CoRR, vol. abs/1810.04805, 2018.

[41] M. Farahani, M. Gharachorloo, M. Farahani, and M. Manthouri, "ParsBERT: Transformer-based Model for Persian Language Understanding," CoRR, vol. abs/2005.12515, 2020.

[42] N. Taghizadeh, E. Doostmohammadi, E. Seifossadat, H. R. Rabiee, and M. Tahaei, "SINA-BERT: A pre-trained Language Model for Analysis of Medical Texts in Persian," CoRR, vol. abs/2104.07613, 2021.

APPENDIX

a) Proposed method pseudo-code

```
Function entityLinking (text, preferLongerEntities, threshold, contextLength): sentences
input text: a piece of text which may contains multiple sentences.
input preferLongerEntities: a boolean.
input threshold: a float number.
input contextLength: a number.
output sentences: tokens of the sentences and their links to the KG.

(allArticleWords, allArticleLinks) = getOrCreateArticleContextCache( )
sentences = createTokens( text )
candidates = generateCandidates( text )
mergeCandidatesToTokens( sentences , candidates, preferLongerEntities )
for ( sentenceIndex , sentence ) in sentences do
context = getContext( sentenceIndex , sentences , contextLength )
context = removeStopWords( context )
candidateGraph = []
for token in sentence do
    for (candidate in token.candidates) do
        candidateGraph.add( allArticleLinks[ candidate ] )
    end for
end for
graphScores= calculateGraphScores( candidateGraph )
for ( tokenIndex , token ) in sentence do
    if token.posTag is (verb or preposition or conjunctions) then
        continue
    end if
    setDefaultMultipliers( token , context )
    setConnectedCandidateMultipliers( token , context )
    for (( candidateIndex , candidate ) in token.candidates) do
        articleWords = allArticleWords[ candidate.title ]
        contextSim = calculateSimilarityOfWords( context , articleWords , token.word )
        graphScore = graphScores[ candidate ]
        assignEntityConfidence( token , candidateIndex , contentSim , graphScore )
        applyFilters( tokne , candidateIndex , context )
    end for
    setBestCandidate( token , threshold )
end for
end for
```

Majid Asgari-Bidhendi is a Ph.D. candidate at Iran University of Science and Technology. He works at IUST Data Mining Lab and researches on open Information Extraction, relation extraction, semantic web, Knowledge Graphs, Entity Recognition and Open Knowledge Extraction.

Farzane Fakhrian received her M.S.c and B.S.c degrees in Artificial Intelligence from the Iran University of Science and Technology in 2018 and 2021. . She is a member of the IUST Data Mining Lab and researches Entity Linking, Machine Learning Methods, and, in particular, Deep Learning in the field of Natural Language Processing.

Behrouz Minaei-Bidgoli is an Associate Professor and the head of the Computer Engineering School at Iran University of Science and Technology. He leads the Data Mining Lab (DML) that does research on various areas in Artificial Intelligence and Data Mining, including Text Mining, Web Information Extraction, and Natural Language Processing. He also acts as the Director of Research for the Department.

[Downloaded from ijict.itrc.ac.ir on 2024-10-12]