Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

A Hybrid Context -Knowledge Representation Model for Arabic Next Word Prediction

Noralhuda N. Alabid

Department of Computer Science, Faculty of Education, University of Kufa, Najaf, Iraq. Email: noralhudan.hadi@uokufa.edu.iq

ARTICLE INFO

Article history:
Received: 17 /11/2025
Rrevised form: 26 /12/2025
Accepted : 04 /01/2026
Available online: 30 /03/2026

Keywords:

Arabic text processing, AraBERT, Knowledge Graph, Language modeling, Next word prediction.

ABSTRACT

One of the controversial topics that has been raised in recent decades. Several approaches involving deep learning and machine learning was implemented to investing in the field of next word prediction in multiple language. In Arabic contexts, this topic is still challenges and in its early stages which need more investigation researches. this field suffers from lakes of robust model that designed specifically for predicted next Arabic word and high-quality dataset that used to develop this domain. This paper proposed a parallel hybrid model (AraBERT- knowledge graph) that augmented the pre-trained AraBERT model with knowledge graph to significantly enhanced next word prediction in Arabic corpus. The constructed of the proposed model involved integrating the AraBERT's contextual vectors with entities of embedding of knowledge graph (KG). The SANAD dataset (195k articles) was used train this model. The model achieved an accuracy of 90% and an F1-score of 91% which outperformed several of fine adaption baseline models including AraGPT-2 (84.8) and AraBERT (82.6%). The significant improvement in results indicates that the combining of contextual and knowledge based has a promising direction for advancing several of Arabic language application including Arabic text prediction, understanding, and auto-generation, and other NLP application.

MSC..

https://doi.org/10.29304/jqcsm.2026.18.12448

1. Introduction

Electronic conversations are a common feature in our modern lifestyle. Mostly, the conversation applications focus on a providing next word prediction system of native language to reduce time spent and facilitate communication. Next word prediction (NWP) is a branch of large language model (LLMs) that concentrate on discovering the most likely word which in a sequence of tokens or words in sentences. This task has contributed to enhancing many applications such as smart authoring and correcting grammatical errors, in addition to more complex tasks such as generating dialogues and machine translation. In addition, this task has applied to enhancing many NLP applications such as smart authoring, correcting grammatical errors, and machine translation.

Despite the importance of the next word prediction, this field remains challenges in Arabic language. It often struggling behind the achieving performance in other languages such as English. This is largely due to the Arabic's linguistic complexities and morphology. A further complicates matters is that the language is divided into three main sections. The formal language, which is based on classical Arabic (the language of the Quran). Modern

∗Corresponding author: Noralhuda N. Alabid

Email addresses: Noralhudan.hadi@uokufa.edu.iq

Communicated by 'sub etitor'

Standard Arabic language, derived from Classical Arabic and used in government, media, and education. The colloquial dialects language, which vary from country to country. These differences make understanding and communication between speakers of different dialects some time difficult.

Deep learning models have mainly contributed to develop words prediction systems, especially when founding the transform-based models such as BERT that makes massive revaluation in this field. It has increased the level of contextual understanding of texts, leading to achieving state-of-the-art results by training on enormous text collections. AraBERT is well known model due to their training on large Arabic corpus and provides improved understanding of linguistic contexts. Despite their power, these models' predictive and inferential abilities depend primarily on the patterns and co-occurrences learned from the corpora. This can cause them to struggle with factual knowledge and long depending knowledge.

To deal with this issues, Knowledge Graphs (KGs) were utilized, which are based on structuring information into entities and defining the relationship between them, as a powerful solution for this limitation. These models provide a clear cognitive representation of real and intuitive reality. The main objective of this research is to integrate the statistical competence gained from AraBERT model with the empirical knowledge of KG models to create a robust and powerful model capable of making predictions based on factual knowledge in addition to statistical analysis. This making system exposes to error when the prediction depends on real word knowledge or logic. The main contributions of this research can be outlined as follow:

1. A novel proposed neural model was suggested that retrieves information from the knowledge graph based on input context and integrates the retrieved information with AraBERT's contextual embeddings.

2. The proposed model was evaluated against several of the AraBERT baselines, which showed significant improvements in prediction accuracy, particularly in texts that required a factual or logical knowledge.

The remainder of this paper as follow: Section 2 reviews the related work in word prediction system. Section 3 describes the proposed methodology for building AraBER-TKG model. Section 4 shows the experimented setting and results, and Section 5 presents conclusion and future direction.

2. Related work

2.1 Next-Word Prediction in Various Languages

Shakhovska et al. [1] suggested applying a hybrid model of LSTM-Markov approach for predicting next words in Ukrainian languages. The LSTM-Markov model achieved an accuracy of75 %. Their results indicated the hybrid model implemented a slow result compared to the performance of Markov when used individually. Sharma et al. [2] applied two deep learning approaches (LSTM and BiLSTM) to predict the next word in a Hindi sequence. They designed a novel model which is based on natural language to facilitate the process of understanding for individuals. The observed results were an accuracy of 59.46% with LSTM and 81.07% with BiLSTM. Atçılı et al. [3] conducted a comparative study between RNN-GRU and LSTM to identify the best approach to predict next word in Turkish dataset. The dataset consisted of 100434 words that obtained from open sources in domain of different sport subjects including football, tennis, volleyball, and basketball. The LSTM model showed the best accuracy results of 81% while RNN-GRU achieved accuracy of 72%. However, in spite of the enhanced results recorded with LSTM, the model was subject-dependent. This means that it could yield different performance with different articles subjects. Singh et al. [4] presented a novel hybrid approach for predicting next words in Punjabi-English language. They applied two deep learning approaches: CNN, for extracting local dependency, and LSTM for detecting long term dependency. The main corpus of the model was collected from Twitter and WhatsApp application, with 500,000 tweets and from 600,000 WhatsApp messages. Their proposed hybrid model achieved an accuracy of 0.93 which was enhanced results compared to other five model implemented on same subject. Al-Anzi et al. [5] conducted ARABERT-LSTM model to predict next word in Arabic contexts. Authors aimed to achieve state of art model on Arabic language by utilizing AraBERT for embedding to enhance the semantic relationship with texts. The ARABERT-LSTM achieved an accuracy of 75% and 64% for LSTM without uses ARABERT. The results showed the improved results in performance that outperform some of the baseline models applied on Arabic texts. The size of these results is restricted with specific scale of training dataset and sources which limiting the batch size, and number of epochs. Tiwari et al. [6] presented neural network architecture of LSTM and BiLSTM to design a next word prediction model suggested model which outperformed several neural network models based on utilized the IITB English–Hindiparallel training dataset. Lahrache et al. [7] investigated both temporal convolutional networks (TCN) and recurrent neural networks (RNN) to determine next words. They used three datasets: Coursera Swiftkey, the book Writings of Friedrich Nietzsche, and the News articles obtained from Brown corpus. The model achieved

an accuracy of 71.51 for RNN and 65.20% for TCN. Hoque et al. [8] utilized n-gram dataset for their proposed model that based on GRU-RNN approaches for next word prediction system in Bangla language. The outcomes displayed notable outcome in accuracy results of 88%, 99%, 97%, and 99% for Unigram model, Bi-gram model Tri-gram, and 4-gram models, respectively. Ikegami et al. [9] presented a hybrid model (RNN-LM) to enhance Japanese text prediction system. The RNN-LM model included an input layer supplied with word embedding, an output layer, and hidden layers connected with LSTMs. The model was trained on 4 million Japanese sentences extracted from Twitter. The experiments indicated that the obtained result was 10% less confusing than traditional models. In addition, this model is included in IME (Input Method Editor) system. Gerz et al. [10] introduced a large-scale LM applied on 50 different languages. The authors offered a novel approach that focused on sub-word-level data which showed notable performance across all 50 languages. Ahmad et al. [11], implemented a comparative study between several of deep learning model to identify the best model used to predict next word system in Urdu languages. They used RNN, LSTM, and BiLSTM on Ur-MonoUrdu dataset containing 8,000 sentences. They conducted their experiments in two phases. In the first phase, they applied the models on a small database (3000 sentences), and the results were: 87%, 79%, and 81% for RNN, LSTM, and BiLSTM, respectively. In the second phase, they expanded the dataset to 5000 sentences to obtain results of 72% for RNN,80% for LSTM,84% for BiLSTM. Mahbub et al. [12] study examined which embedding method was best suited for integration with LSTM to predict next word in Bengali corpus of 25,000 sentences. They used word2vec skip-gram, word2vec CBOW, fast-skip-gram and fast-text-CBOW. Results exhibited that word2vec-gram achieved the best accuracy of 79.72% compared to others. Shahid et al. [13] proposed deep learning model for Urdu next word prediction. They investigated LSTM and BERT on large Urdu dataset contain of one 1.1 million terms for training both models. The accuracy scored of 52 for LSTM and 74 % for BERT. Other researchers [14] have also evaluated the role of LSTM in next words prediction but with the Assamese texts (the language of local India). Their model achieved a promising result which demonstrating a high level of accuracy of 88%. Badawi [15] incorporated N-gram model for next word prediction in Kurdish language. Despite the lack of large Kurdish dataset, as well as the insufficient number of N-gram models for the Kurdish language, their model yielded a noticeable result 96% for accuracy.

2.2 Knowledge-Enhanced Language Models

Several of resent studies were motivated to move from sequential research models to explore structured knowledge in exploring large language models. Integrating the KGs in LLMs is valuable for obtaining accurate results with answer question models [16]. An example of this is the ReLMKG model [17], which employed the KG for answering the complex question based on expanded information sources. Zhang et al. [18] proposed KnowGPT, a framework model designed to enhance the accuracy of LLMs tasks by involving the factual knowledge (KG) model into their response. This model achieved 92 % accuracy in openbookQA task. Their model has shown the potential importance of using KG model to extract the relevant knowledge and convert it into effective factual prompts. However, this model suffered from complex structure that makes the implementation of this model require high computational resources. Zhang1 et al. [19] proposed a novel model (GLAME) that utilizes a pre-trained LLMs with knowledge graph (KG) to detect the relevant knowledge and inject its structure into LLMs. Despite the improvements achieved (6%) over prior models, its performance depends primarily on the presence and quality of the underlying knowledge graph. This makes the model's accuracy dependent on an external knowledge source. For task like next word prediction, Magar et al. [20] suggested Graph Neural Network combined with LSTM to encode the local context of preceding words and predict next words. They used Wikipedia dataset of 1.9 billion word including subjects of music, sports and celebrities. They obtained an accuracy of 53%,61%, 67% for sports, celebrities, and music, respectively. Although these methods have proven the importance of involving knowledge integration with LLMs, it has not yet been exploited in special tasks such as next word prediction in the Arabic language. the proposed model AraBERT-KG model addresses this gap by designing a hybrid model that performes a parallel task to leverage the power of structured knowledge for Arabic next word prediction.

3. Methodology

3.1 Dataset description and preprocessing

The SANAD dataset is used in experiments. It is an Arabic dataset [21] collected from three news resources: AlKhaleej, AlArabiya, and Akhbarona. The dataset has 194,797 Arabic articles categorized into seven domains: Culture, Finance, Medical, Politics, Religion, Sports, and Tech. each article in the SANAD dataset is single label, meaning that each article is assigned to one specific domain. Table 1 shows the statistical distribution of the dataset.

Table 1. The statistic distribution of the SANAD dataset.

Dataset	No. Articles
AlKhaleej	78050
AlArabiya	71247
Akhbarona	45500
Total	194.797

The articles of the dataset were organized into seven folders corresponding to their perspective domains. Each article is saved in text file of (.txt format). For processing, simple python scripts were implemented to read the data and prepared it for training the proposed model. The dataset was randomly split into two sub-files: training and testing subsets with a range of 80:20. The training file was split into 90% located for training phase and 10% for validation step. Table 2 displays the count of training, validation and testing of the SANAD dataset. For preparing the corpus for processing, several of tradition preprocessing methods were used such as removing white space, punctuation marks, and URL.

Table 2. Size of train, test, and validation subsets.

Dataset	Train file	Test file	Validation file
SANAD dataset	140.254	38.959	15.584

3.2 Knowledge graph construction

Knowledge Graph (KG), which is also named as Knowledge Base, is a structured representation that based on organizing real word entities and their interlinking in a graphlike structure. Typically, it helps model unstructured information in a significant graph [22]. In the graph, nodes represent entities and edges between them defines the type of relationship. In addition, each node can share its information with other nodes. It has been an essential resource for many NLP applications that rely on knowledge such as Question Answering system [23], medical industry [24], and e-commerce [25]. KG is defined by a triple of (subject, predicate, object). Fig. 1 shows an example of a KG model illustrating the triple of ("النصر"," فاز على"," الاتحاد ") ("النصر"," ينتمي الى "," الرياض ").

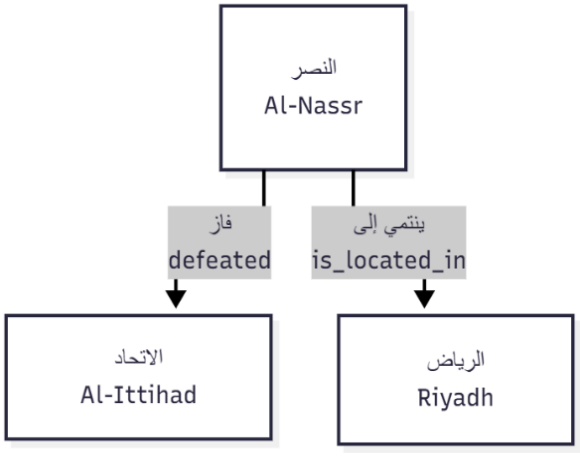

Fig. 1 The illustration entities with KG.

3.3 Feature extraction using AraBERT

Original input texts cannot be fed directly into deep learning models. They must first be transformed into numerical vectors by using embedding techniques. In this proposed model, AraBERT was selected to generate the contextual feature embeddings from the input corpus based on linguistic analysis, while the Knowledge graph operates in parallel to retrieve at the same time the KG work on retrieving entities embedding based on knowledge

understanding. Basically, AraBERT is based on the significant BERT architecture that uses a transformer for embedding to capture the meaning of words in contexts. Furthermore, AraBERT is distinguished by its ability to capture the syntactic and semantic relationship with keeping the nuanced meaning of each word within the related sentence. In this system, BERT-large-arabertv02 was used [26]. At first, the input strings were tokenized using the AraBERT tokenizer. After that, the tokenized sequences were processed through the pre-trained AraBERT model. The output of the last hidden layer of AraBERT model were extracted, and mean pooling was applied across all token location. This creates an embedding vector of 768 dimension for each input sequence. These encoder vectors form the foundation for creating the knowledge schema and determining the final word prediction.

3.4 Prediction model with KG

The main aim of the proposed system is to design a hybrid neural model that integrates contextual embeddings from the AraBERT with a knowledge graph (KG) to predict the next word. Figure 2 illustrates the proposed model. The input strings were tokenized and processed by using the BERT-large-AraBERT model. At the same time, the interesting entities of the input strings were identified by matching words against the entity set of the knowledge graph. The 2-layer Graph attention network (GAT) was used to capture the relationship between these entities. Then, the precomputed feature vector for each identified entity was retrieved. Node2Vec model was used to generate 64-dimensional embedding vector on SANAD knowledge graph. These vectors included the encoding information which is specific for each node such as degree, weighted connectivity, and its domain (e.g., sport, politics). The entity embeddings for a given input were averaged and projected to a space of 64 dimensions. The 768-dimensional AraBERT vector was combined with 64-dimensional KG vector resulting in 832-dimensional representations. This combined vector was passed across a fusion layer (fully connected layer with ReLU activation) for the next word prediction. The output of fusion layer was fed into a final linear layer. This layer was responsible for producing the probability distribution matrices that define the final next word prediction.

3.5 Experiential setup and training

The model was implemented using PyTorch and Transformers library. As previously mentiond, the pre-trained BERT-large-AraBERTV2 model from Hugging Face Hub website was used. The hyperparameters were set empirically to obtain the best results. The model was trained with AdamW optimizer using the following configuration: a learning rate of 0.00002, and weight_decay of 1e-4. A StepLR scheduler was applied to reduce the learning rate by a value of 0.8 every two epochs. The Cross Entropy loss function was used. Training was implemented for 5 epochs using a batch size of 8.

3.6 Evaluation matrix

To evaluate the proposed model, the standard evaluation matrixes that used in Natural language processing (NLP) which are: accuracy, precession, recall, and F1-score. These matrixes are consisted from several values including True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN)) which are extracted form confusion matrix. The result of accuracy is obtained by divided the ration of correctly detected phrase on the overall number of evaluated ones. The recall value is determined by calculating the percentage of total correct found phrases. Precision represents the rate of correct samples divided by the total number of correct predicted phrase by the model. F1-score is founded by calculating the mean of precision and recall values. They are calculated based on follow formulas.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

$$\text{F1-Score} = 2 * \frac{precision*recall}{precision+recall} \tag{4}$$

Table 3. Explain the parameter configuration of the proposed model.

Models	Parameters	Used values
AraBERT	The basic mode	BERT-large-AraBERTV2
	Hidden size	768
	Number of layers	12
	Pooling method	Mean pooling
KG	KG dimension	64
	Number of layers	3
Fusion layer	Final size of vector	832
	Fusion method	ReLU function
Training setting	Optimizer	AdamW
	Learning Rate	2e-5
	weight_decay	1e-4
	learning rate scheduler	StepLR (gamma=0.8, step size=2 epochs
	Batch size	8
	Number of epochs	5

4. Results and discussions

4.1 Results

The AraBERT embedding method was used in a model combined with a knowledge graph to predict next word in Arabic strings. The model was trained for 20 epochs on SANAD dataset of 194.797 articles. Table 4 displays the enhanced results obtained when combining AraBERT with Knowledge graph model. The proposed model (AraBERT-KG) achieved an accuracy of 90% and F1-score of 91% while applied the Knowledge graph to predict next word yielded an accuracy of 82%, and a F1-score of 80%s. Table 5 illustrates sample sentences that used to evaluate the system.

Table 4. The results of the proposed model.

Model	Accuracy	F1-score	Precision	Recall
Knowledge graph (KG)	84	81	80	83
AraBERT-KG	90	91	88.6	93

Table 5. Shows some examples for predicting next word by the proposed model.

Sentences	The sentence After replace word with mask	Predicted index	Predicted word
مشروع الحملة الوطنية للتخلص من الالتهاب الكبدي	مشروع الحملة الوطنية للتخلص من الالتهاب[MASK]	8724	الكبدي
إن حضور المهرجان جسد جواً من حسن الحوار الثقافي	إن حضور المهرجان جسد جواً من حسن الحوار [mask]	**3056**	الثقافي
ان هدفنا هو تعريف المعنيين بحماية الطفل	ان هدفنا هو تعريف المعنيين بحماية[MASK]	5428	الطفل
كشفت صحيفة ذي صن البريطانية الشعبية	كشفت صحيفة ذي صن البريطانية[MASK]	6542	الشعبية
اتساع السوق خاصة مع زيادة عدد حاملي الهواتف الذكية	اتساع السوق خاصة مع زيادة عدد حاملي الهواتف [mask]	9843	الذكية
مجموعة متنوعة من الأجهزة المنزلية	مجموعة متنوعة من الأجهزة[Mask]	11045	المنزلية

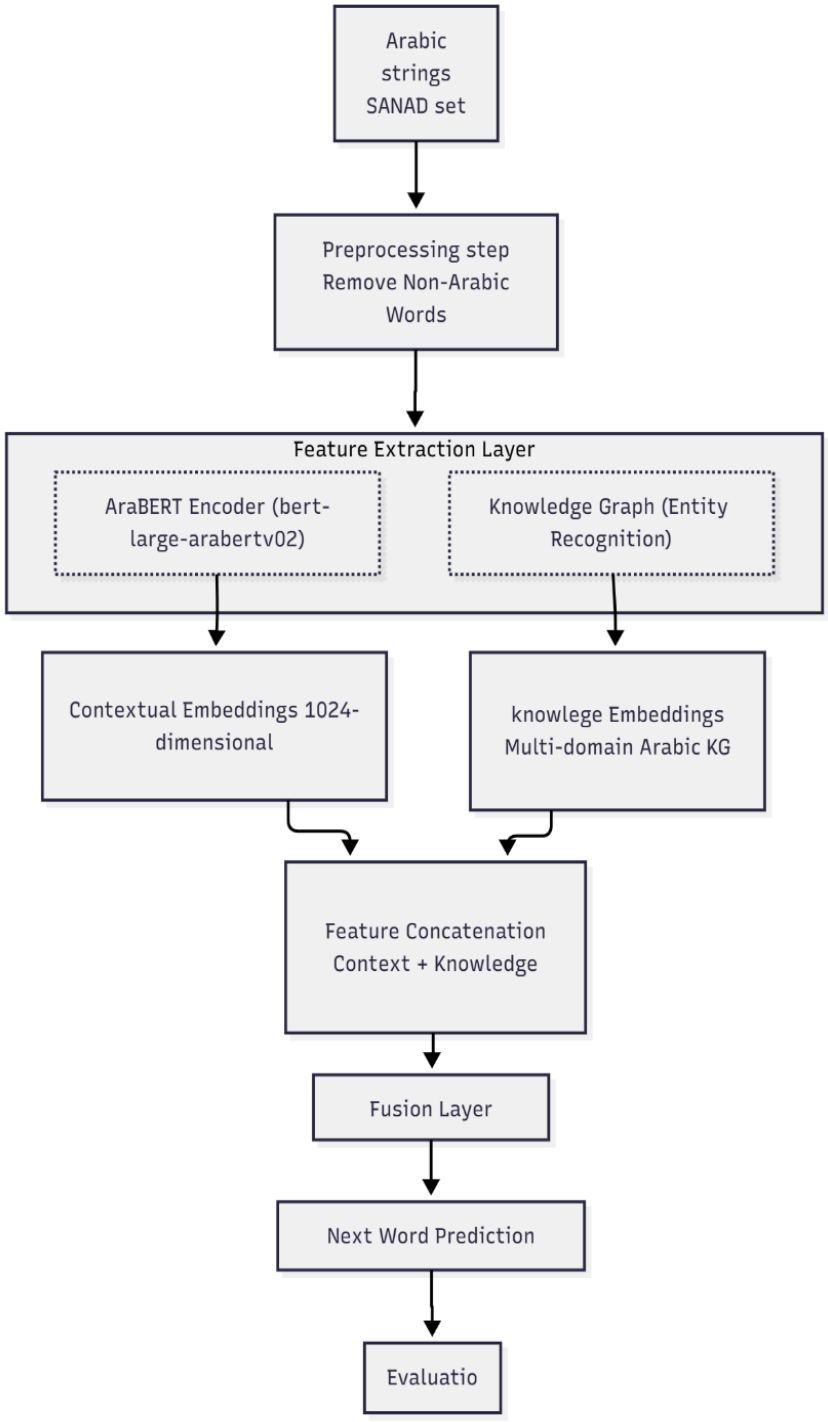

Fig. 2 The architecture of the proposed model for predict next word.

4.2 Discussions

The proposed model demonstrated a relative improvement (6%) in an accuracy compared to Knowledge graph-based model. Furthermore, AraBERT-KG model achieved enhanced results in F1-score which indicates an improvement in balancing between precision and recall. The integration of AraBERT with Knowledge graph proved effective in handling contextual semantics in next word prediction system for Arabic texts. The system successfully dealt with the complexities of the linguistic structure in Arabic language.

However, the error rate of (10%) could be a result of complex morphological structures in some Arabic words. In addition, some of Arabic sentences could be ambiguous and offer more than one option in prediction process. To investigate this error, subset of incorrect model prediction is illustrated in Table 6.

Table 7 shows some existing methods which were used for comparison with the proposed model in order to assess the performance of the system. However, to make reasonable and strong comparable baseline, these baselines models were redeveloped and tested them by utilizing my benchmark for Arabic next word prediction. Antoun et al. used AraGPT-2 for Arabic language generation. They used large collection of online Arabic texts for adaption a powerful architecture model for Arabic language modeling [27]. By applying their model on SANAD dataset, I was able to determine its performance on my benchmark and compare with the proposed model. The AraBERT Language understanding model [26] was redeveloped to predict Arabic next word by SANAD training set. The author of [5] used AraBERT combined with LSTM for predicting Arabic next word. Their dataset was collected by converting Arabic audio data into text of size 1,925,826 sentences.

Table 6. Illustrates the failure case of the proposed model.

Detected error	Example	Possible reasons
complex morphological	...نحو 20 فناناً تشكيلياً من مختلف ..... العالم، إلى جانب... Error detection:دولة (county) Correct word : دول (countries)	The model failed to achieve agreement between numerical phrases and pronouns.
Failure in KG model	....تتوقف محطة الوقود بسبب ..... Error detection:الاحتفال (celebration) Correct word : نقص الامدادات (lack of supplies)	KG model fails to detect the most relevant phrase
Complex morphological	....بادر بالانضمام إلى قائمة الشركاء الاستراتيجيين للهيئة، ممن في أعمال الخير. ..... Error detection:المتعاونون (collaborators) Correct word: يتعاونون(cooperate)	Failure in detecting the correct morphological derivation word
Differences in dialects	...واختتم العويس تصريحه قائلاً .... على يقين أن النجاح سيكون حليف الأوفياء .... Error detection:أنا (I) Correct word: نحن (we)	The model failed to adapt with the formal language of the news.

Table 7. Displays the performance comparison between the proposed AreBERT-KG model against other existing Arabic next word prediction methods on SANAD dataset.

Model	Implementation detail	Accuracy
raGPT-2	The model in [27] was adapted for next word prediction by utilizing SANAD corpus. **Original task**: Arabic language generation	84.5%
ARABERT	The pre-train model in [26] was fine-tuned using the SANAD dataset for Arabic next word prediction. **Original task:** Arabic language understanding	82.6%
LSTM	The standard LSTM model in [5], which was designed for Arabic next word prediction, was adjusted on the SANAD dataset. The original model achieved 64.9% Accuracy from audio derived dataset.	69.3%
ARABERT + LSTM	The hybrid model in [5] (design for Arabic next word prediction) was re-implemented on the SANAD dataset. The original model yielded 74.6% accuracy based on audio derived dataset	77.8%
Proposed model (ARABERT-KG)	-	90%

5. Conclusion

This work bridges the gap between statistical understanding of language model and structured cognitive illustrated for next word prediction in Arabic language. The proposed model combines AraBERT (to generate contextual features vectors based on linguistic analysis) with a knowledge graph (to provide factual knowledge). The hybrid model AraBERT-KG was trained on the SANAD dataset which contain 194, 797 articles. The results indicate the importance of considering statistical representation in addition to cognitive representation in the process of determining the next word. several of previous model were fine turned on the benchmark model within experimental structure to create a fair and direct comparison. These models were investigated in different areas of LLMs models. This is because of lack of previous Arabic research using knowledge graph (KG) for Arabic next word prediction, addition to the limited number of studies applied to the SANAD dataset. The proposed system obtained an accuracy of 90% and F1-score of 91% which is much better than enhanced baseline model. This system opens new opportunities for developing several NLP models which based on knowledge enhanced representation in Arabic language such as machine translations, text similarity, and question answering applications. Despite the promising results, error analysis indicates that the system struggles to distinguish ambiguous contexts that may have more than one correct prediction. These difficulties could likely be overcome by employing richer contextual inference. In addition, for future, it is aimed to expanding the knowledge graph representing to covered more domain entities and relationship. It is also possible to adapt the system to investigate more Arabic dialects.

Conflicts of interest

The author has no conflicts of interest to declare.

Data availability

The SANAD dataset used in this system is online available and can be download from

References

[1] K. Shakhovska, I. Dumyn, N. Kryvinska, and M. K. Kagita, "An Approach for a Next-Word Prediction for Ukrainian Language," *Wirel. Commun. Mob. Comput.*, vol. 2021, no. 1, Jan. 2021, doi: 10.1155/2021/5886119.

[2] R. Sharma, N. Goel, N. Aggarwal, P. Kaur, and C. Prakash, "Next Word Prediction in Hindi Using Deep Learning Techniques," in *2019 International Conference on Data Science and Engineering (ICDSE)*, IEEE, Sep. 2019, pp. 55–60. doi: 10.1109/ICDSE47409.2019.8971796.

[3] A. Atçılı, O. Özkaraca, G. Sarıman, and B. Patrut, "Next Word Prediction with Deep Learning Models," in *Smart Applications with Advanced Machine Learning and Human-Centred Problem Design*, 2023, pp. 523–531. doi: 10.1007/978-3-031-09753-9_38.

[4] G. Singh and C. P. Kamboj, "Deep Learning for Predicting the Next Word in Bilingual Social Media Texts," *SN Comput. Sci.*, vol. 6, no. 54, 2024, doi: doi.org/10.1007/s42979-024-03585-8.

[5] F. S. Al-Anzi and S. T. B. Shalini, "Revealing the Next Word and Character in Arabic: An Effective Blend of Long Short-Term Memory Networks and ARABERT," *Appl. Sci.*, vol. 14, no. 22, p. 10498, Nov. 2024, doi: 10.3390/app142210498.

[6] A. Tiwari, N. Sengar, and V. Yadav, "Next Word Prediction Using Deep Learning," in *2022 IEEE Global Conference on Computing, Power and Communication Technologies (GlobConPT)*, IEEE, Sep. 2022, pp. 1–6. doi: 10.1109/GlobConPT57482.2022.9938153.

[7] F. LAHRACHE and S. DJEBRIT, "Next Word Prediction Based On De ep Learning," University of Ghardaia, 2020. [Online]. Available: https://dspace.univ-ghardaia.edu.dz/xmlui/handle/123456789/366

[8] A. Hoque, B. Jahan, S. C. Paul, Z. A. Zabu, R. Mondal, and P. Akter, "Next Words Prediction and Sentence Completion in Bangla Language Using GRU-Based RNN on N-Gram Language Model," *J. Data Anal. Inf. Process.*, vol. 11, no. 04, pp. 388–399, 2023, doi: 10.4236/jdaip.2023.114020.

[9] Y. Ikegami, S. Tsuruta, A. Kutics, E. Damiani, and R. Knauf, "Fast ML-based next-word prediction for hybrid languages," *Internet of Things*, vol. 25, p. 101064, Apr. 2024, doi: 10.1016/j.iot.2024.101064.

[10] D. Gerz, I. Vulić, E. Ponti, J. Naradowsky, R. Reichart, and A. Korhonen, "Language Modeling for Morphologically Rich Languages: Character-Aware Modeling for Word-Level Prediction," *Trans. Assoc. Comput. Linguist.*, vol. 6, pp. 451–465, Dec. 2018, doi: 10.1162/tacl_a_00032.

[11] M. H. Ahmad, A. Saeed, M. U. Bhatti, N. Hussain, M. F. Ullah, and M. Anwar, "Next Word Prediction for Urdu using Deep Learning Techniques," *VFAST Trans. Softw. Eng.*, vol. 13, no. 1, pp. 49–59, Feb. 2025, doi: 10.21015/vtse.v13i1.2044.

[12] M. Mahbub, S. Akhter, A. Kabir, and Z. Begum, "Context-based Bengali Next Word Prediction: A Comparative Study of Different Embedding Methods," *Dhaka Univ. J. Appl. Sci. Eng.*, vol. 7, no. 2, 2022, doi: https://www.banglajol.info/index.php/DUJASE/article/view/65088/44180.

[13] R. Shahid, A. Wali, and M. Bashir, "Next word prediction for Urdu language using deep learning models," *Comput. Speech Lang.*, vol. 87, p. 101635, Aug. 2024, doi: 10.1016/j.csl.2024.101635.

[14] P. P. Barman and A. Boruah, "A RNN based Approach for next word prediction in Assamese Phonetic Transcription," *Procedia Comput. Sci.*, vol. 143, pp. 117–123, 2018, doi: 10.1016/j.procs.2018.10.359.

[15] H. K. Hamarashid, S. A. Saeed, and T. A. Rashid, "Next word prediction based on the N-gram model for Kurdish Sorani and Kurmanji," *Neural Comput. Appl.*, vol. 33, no. 9, pp. 4547–4566, May 2021, doi: 10.1007/s00521-020-05245-3.

[16] X. Liang, Z. Wang, M. Li, and Z. Yan, "A survey of LLM-augmented knowledge graph construction and application in complex product design," *Procedia CIRP*, vol. 128, pp. 870–875, 2024, doi: 10.1016/j.procir.2024.07.069.

[17] X. Cao and Y. Liu, "ReLMKG: reasoning with pre-trained language models and knowledge graphs for complex question answering," *Appl. Intell.*, vol. 53, no. 10, pp. 12032–12046, May 2023, doi: 10.1007/s10489-022-04123-w.

[18] H. Chen, J. Dong, X. Huang, Z. Yu, D. Zha, and Q. Zhang, "KnowGPT: Knowledge Graph based Prompting for Large Language Models," in

Advances in Neural Information Processing Systems 37, San Diego, California, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024, pp. 6052–6080. doi: 10.52202/079017-0196.

[19] M. Zhang, X. Ye, Q. Liu, P. Ren, S. Wu, and Z. Chen, "Knowledge Graph Enhanced Large Language Model Editing," *arXiv Prepr. arXiv2402.13593*, 2024, doi: https://doi.org/10.48550/arXiv.2402.13593.

[20] A. T. Magar and A. Shakya, "Next Word Suggestion using Graph Neural Network," *arXiv:2505.09649v1*, 2025, doi: doi.org/10.48550/arXiv.2505.09649.

[21] O. Einea, A. Elnagar, and R. Al-Debsi, "SANAD: Single-Label Arabic News Articles Dataset for Automatic Text Categorization," *Data Br.*, vol. 25, no. 2, 2019, doi: 10.17632/57zpx667y9.2.

[22] W. Wu, C. Wen, Q. Yuan, Q. Chen, and Y. Cao, "Construction and application of knowledge graph for construction accidents based on deep learning," *Eng. Constr. Archit. Manag.*, vol. 32, no. 2, pp. 1097–1121, Feb. 2025, doi: 10.1108/ECAM-03-2023-0255.

[23] M. Dehghan *et al.*, "EWEK-QA : Enhanced Web and Efficient Knowledge Graph Retrieval for Citation-based Question Answering Systems," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 14169–14187. doi: 10.18653/v1/2024.acl-long.764.

[24] M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, and D. Sontag, "Learning a Health Knowledge Graph from Electronic Medical Records," *Sci. Rep.*, vol. 7, no. 1, p. 5994, Jul. 2017, doi: 10.1038/s41598-017-05778-z.

[25] K. K. Teru, E. G. Denis, and W. L. Hamilton, "Inductive relation prediction by subgraph reasoning," *37th Int. Conf. Mach. Learn. ICML 2020*, vol. PartF16814, no. 1, pp. 9390–9399, 2020.

[26] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," in *LREC 2020 Workshop Language Resources and Evaluation Conference 11--16 May 2020*, 2020, p. 9.

[27] W. Antoun, F. Baly, and H. Hajj, "ARAGPT2: Pre-Trained Transformer for Arabic Language Generation," *Proc. sixth Arab. Nat. Lang. Process. Work.*, pp. 196–207, 2021, doi: arXiv:2012.15520.