Survey: using BERT model for Arabic Question Answering System.

Abdullah Farhan Mahdi
University of Diyala, Department of Computer/Baqubah, Iraq
abd.farhan71@gmail.com

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 4 June 2021

Abstract— In this paper, we deal with the community question answer problem. Using Burt's algorithm, the Question Answer Task (QA) is a Natural Domain Language Processing (NLP). We present a survey on language representation learning for the purpose of consolidating a set of common lessons learned across a variety of recent efforts. Which enables machine reading comprehension and natural language inference tasks. BERT controls its simplicity of use also is a light refinement method without substantial task-specific modifications. We highlight important considerations when interpreting recent contributions and choosing which model to use. We will address the strengths and weaknesses of the algorithm and what are the challenges that faced researchers.

I. INTRODUCTION

Question answering systems(QA) are the most important areas in natural language processing(NLP). that rely on retrieval of complete answers[1]. The questions are classified into three main types, yes / no, why, definition, each type has a different approach[6][7]. In the answer modeling problem, the system for answering questions in all three forms is a fil Language template that takes a question and its context document It results in the answer to the question, given the content. Give a yes / no question (short The title can give a great description) and a list of the community Answers, decide whether the global answer On the question must be yes, no, or not sure. The possibility of enriching the semantics of the contextual sentence With excellent results specific to the predicate by Introducing SemBERT: semantic-aware BERT file It arranges BERT with well-defined contextual clues. In recent years, at a time of the assisted growth of web use Motivate users to access information and pay attention to quality assurance systems[8]. Semantic matching is important Part of (NLP), It is used to measure the similarities and what is the relationship between the differences for the textual elements , Such as words, sentences, or documents. There are a lot of training samples located within the database dedicated to training in different languages.

Table 1 shows samples from the training dataset.

question1	question2	is_duplicate
ما هي الطرق الصحيحة للاعتناء بالحامل؟	كيف اهتم بطفلي؟	0
ما هي وسائل الاتصالات الحديثة؟	ماذا نعني بوسائل الاتصال الحديثة؟	1
ما طريقة تحضير محشي الكوسا ؟	من طرق تحضير محشي الكوسا؟	1
ما طريقة تحضير حلى الطبقات؟	من طرق تحضير طبقات الكيك؟	0
من الآيات القرآنية عن الراعي والرعية ؟	ما هو تعريف الراعي والرعية ؟	0
أين تقع قارة أوروبا؟	ما هو موقع اليمن؟	0

II. BACKGROUNG

Most of the researches were far from the Arabic language for many reasons, including the complexity of grammar, the language, the lack of vowels, and they do not deal with capital letters despite the presence of more than 25 countries that speak the Arabic language and in the continents of Asia and Africa. There are researches that have developed the answer in the Arabic language .Here are some of that research. Question Answering system to support Arabic language (QARAB) [2] is a system that uses both information retrieval and natural language processing techniques. It has main drawback which is not considering to understand the question in deep semantic level. AQAS [3] is used to answer knowledge-based questions The system does not use the raw data used by the regulator Instead of data. System for answering Arabic questions (Arabic

QA) [4], [5] It is an Arabic question and answer system that mainly relies on the syllable for the purpose of retrieving information

1-BERT Model

Demonstrated pre-training language models Of great importance in improving the performance of many NLP tasks, including Questions of a dual nature[9]. There are two ways to implement the algorithm Pre-trained language presentations on NLP tasks; Either on merits basis or refinement. To Feature based approach[10]. There are limitations to learning in the generic language representations that depend The structures are one-way from left to right. The encryption is bidirectional Transformers from transformers (PERT) It strongly excels in many of its characteristics over the latest previous technology One-way models[11]. This model relies on self-intelligence .Which contains more than one head Accuracy is the mechanism that this algorithm works for and contains multiple tasks such as the question, the answer, and the classification of all the sentences. Cutting edge precision over a wide range of Tasks such as natural linguistic reasoning and question Answer and sentence classification[12]. Each word entered the system learns it through Bi-directional encoder representations using Language form, which is randomly hidden Some words from the input to predict[13]. The left and right shapes are BERT, Encoder and decoder respectively as shown1 in the figure below.

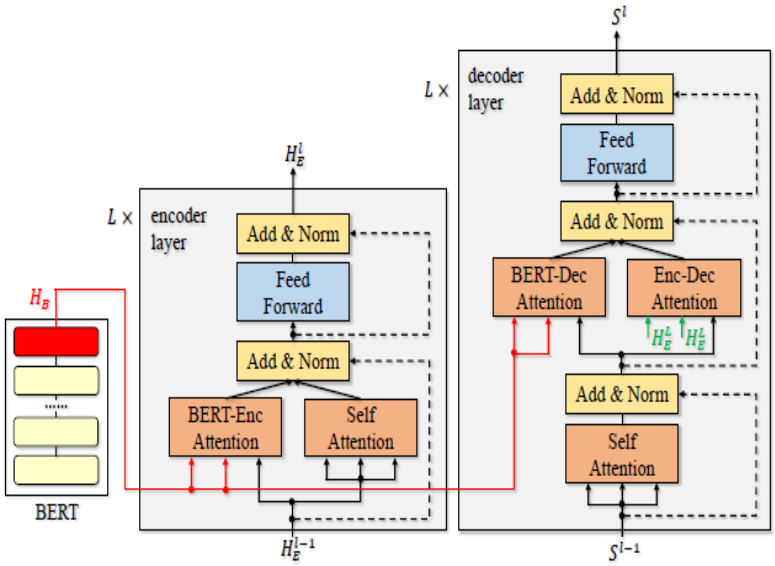

Figure 1: The architecture of BERT model.

In the latest research, with an increase in existing transformers Depending on the language BERT in understanding the language is considered very advanced, it is required to be previously trained on a very large group. These results are capable of achieving NLP tasks[14].

2-Challenges of processing text

There are many challenges as follows:

- Normal language handling has to adapt to the type of text it deals with. Is it a governmental, administrative, or scientific text or a text for the general public also the difference of people in writing the text affects the type of text[15].
- Among the challenges facing text processing are the various social media platforms that contain many different techniques, especially in the aspect of sentiment analysis[16].
- There are texts that have unique characteristics such as the length of the short text, where the number of characters for this text is more than 290 characters[17].
- There are limitations and restrictions on the length of the tweet, forcing the user to find a shortcut, which causes them to not understand the true meaning of the text[18].
- Some texts contain irregular content of the language words, misspelled words and colloquial words.

- Spelling mistakes distract the data, which in turn affects the context of the sentence system[91].
- Feelings of longing and loss because of death .They give a positive and sympathetic meaning, for example, God will have mercy on you and make your rest, paradise in general is positive, but the feeling here is death[20][21].
- One of the greatest challenges is building a language that is extracted from resources because of the great difference that exists between one language, including eloquent and colloquial ones. Rather, it came that one language is spoken by humans, each region differently from the other[22].

3-Related Works

In [23] the researcher produced the solving of the problem concerning the open domain factual Arabic question answering (QA), by using the Wikipedia knowledge source. their system consists of a step of document retriever using the TF-IDF and the using BERT for document reader. The results showed the effectiveness of translating data as a training resource for QA.

In[24], the researcher utilizing BERT by pre-trained it especially for the Arabic language, and comparing the performance of AraBERT to a multilingual BERT. The achieving results proved the effectiveness of the developed AraBERT in the state of the art performance on most tested Arabic NLP tasks.

In [25], the researcher produced a methodology for creating the annotation process and corpus and developing two machine learning baselines for two designed tasks: stance prediction and claim verification, by employing best model utilizes pre-training (BERT) and achieves 76.7 F1 on the stance prediction task and 64.3 F1 on the claim verification task. Results hint that while the linguistic features and world knowledge learned during pre-training are useful for stance prediction, such learned representations from pre-training are insufficient for verifying claims without access to context or evidence.

In[26], the researchers reported a model for detecting special information such as age, gender, and language variety from the user's social media data in the context of the Arabic author profiling and deception detection shared task (APDA). They build simple models based on pre-trained bidirectional encoders from Transformers (BERT). They acquire 54.72% accuracy for age, 93.75% for dialect, 81.67% for gender, and 40.97% joint accuracy across the three tasks.

In[27], this work showed that a simple neural model using multilingual BERT had a competitive performance that is superior to traditional classifiers that use many hand-crafted features for the task. The results proved that the use of a language model pre-trained on Arabic data only can yield better performance and thus, then planning to experiment with such models next. The researchers hypothesize that including some of the hand-crafted features in the neural model can bring improvements to the performance and then they plan to test this hypothesis in future work.

In[28], the paper proposes a semi-supervised learning approach to train a BERT-based NER model using labeled and semi-labeled datasets. paper presented a new approach to detect and classify named entities in any Arabic text. Their approach consists of training an already pre-trained BERT model for Arabic NER in a semi-supervised fashion.

In[29], the researcher aims to discuss the current state-of-the-art and remaining challenges, outline requirements and suggestions for practical parallel data collection, and describe existing methods, benchmarks, and datasets. Then demonstrate that a simple translation of texts can be inadequate in the case of Arabic, English, and German languages (on InsuranceQA and SemEval datasets).

In [30], the researcher uses state-of-the-art transformer models to train the QA system on a synthetic reading comprehension dataset translated from one of the most popular benchmark datasets in English called SQuAD 2.0. Then collecting a smaller human-annotated QA dataset from Bengali Wikipedia with popular topics from Bangladeshi culture for evaluating our models. Finally, they compare their models with human children to set up a benchmark score using survey experiments.

In[31], this paper produced the uses of the deep learning approach to tackle the Arabic NER task. The researcher introduced a neural network architecture based on bidirectional Long Short-Term Memory (LSTM) and Conditional Random Fields (CRF) and experimented with various commonly used hyper parameters to assess their effect on the overall performance of our system. The model gets two sources of information about words as input: pre-trained word embedding's and character-based representations and eliminated the need for any task-specific knowledge or feature engineering. They obtained state-of-the-art results on the standard ANERcorp corpus with an F1 score of 90.6%.

4- Analysis and Discussion

A comparison of Burt's work in the question-answering system

The main objective of this literary survey is to study the different works that are related to the same topic and to show how methods of work in answering questions for the same algorithm in different ways and in different languages.

Team	**Preprocessing**	**Methods**	**Weakness**	**Effectivenesss**
(Djandji et al., 2020)	He removed all non-Arabic letters and fragmentation from the words	Controlling and accurate weddings according to context(Arabart)and eliminating perversion	The researcher used the source of knowledge specific to one topic for the reader of any document	Translate all data for the purpose of ensuring high quality
(Elmadany et al., 2020)	Names, numbers, and all tags and associations replace with NUM, USER, HASH and URL respectively	Use multilingual, sentimental, BERT-based forms	The researcher did not address any of the questions that are the most accurate	The researcher used BERT's algorithm for the purpose of comparison and got a high efficiency
(Saeed et al., 2020)	Using regular letters, removing any repeated letters, dividing them into words	(Multilingual BERT) and non-contextual weddings Aravec , Fast Text, word2vec) with classifier group that compiled output using SVM,RF, NB, etc.	Representations gained from prior training are insufficient to verify claims without access to context or evidence.	Two designed tasks: situation prediction and claim verification
(Hassan et al., 2020)	diacritic, kashida, repeated letter, and non-Arabic character removal	Pre-trained character set by multi-language BERT	They build simple models based on pre-trained	They acquire 54.72% accuracy for age, 93.75% for dialect, 81.67% for gender, and 40.97% joint accuracy across the three tasks.
Keleg et al., 2020)	word segmentation	(multilingual BERT and AraBert) contextual embeddings	Some hand-made features delay the work and lead to slow results	use of a language model pre-trained on Arabic data only can yield better performance
Mozannar et al., (2019)	factual Arabic question answering (QA), by using the Wikipedia	step of document retriever using the TF-IDF ,the using BERT for document reader.	Training is almost subject to supervision, its results are stable	Use groups of classified and semi-labeled data.
Khouja, J. (2020).	Demonstration learning for machine learning with two lines of position prediction and verification of claims	Linguistic features and global knowledge learned during previous training are useful for predicting the situation, uses pre-training (BERT)	translation of texts can be inadequate in the case of Arabic, English, and German	discuss the current state-of-the-art and remaining challenges

Zhang, C., et al., (2019).	Model for detecting special information such as age, gender, and language variety from the user's social media data	simple models based on pre-trained bidirectional encoders from (BERT).Transformers	The use of human experiment models	The researcher uses the latest transformer models to train the quality assurance system on a set of data
Hasanain, Hasanain, ., et al., (2020).	The use of a language model pre-trained on Arabic data only, can yield better performance	simple neural model ,using multilingual BERT	Rely on bi-directional long-term memory (LSTM) and random fields	They obtained state-of-the-art results on the standard ANERcorp corpus with an F1 score of 90.6%.

Table 2: Different methods used by different teams

5- Conclusion

This analysis dealt with most of the applications that are interested in answering the questions that used the PERT algorithm. The main objective of this survey is to present the methods used by most researchers and to assess the repercussions of repetition in the methods and methods used. Eleven studies have been published in recent years, which have kept pace with the development in answering questions. This survey proved that most of the users of the PERT algorithm rely on prior training of the data to obtain good results. Lately, most work has focused on the three methods of question and answer.

6-References

[1] J. Rose Finkel, T.Grenager, and C. Manning. 2005." Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling". Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370.

[2] E. Breck, J. Burger, D. House, M. Light , I. Mani (1999) ``Question answering from large document collections'', Question Answering Systems: Papers from the 1999 AAAI Fall Symposium, 5-7 November, North Falmouth, MA, AAAI Press, Menlo Park, CA, pp. 26-31.

[3] F.A. Mohammed, K. Nasser, H.M. Harb (1993), "A knowledge-based Arabic Question Answering System (AQAS)". In: ACM SIGARTBulletin, pp. 21-33.

[4] Y. Benajiba, P. Rosso, A. Lyhyaoui, 2007. "Implementation of the ArabiQA Question Answering System's components". In: Proc.Workshop on Arabic Natural Language Processing, 2nd Information Communication Technologies Int. Symposium, ICTIS- 2007, Fez,Morroco, April 3-5.

[5] Y. Benajiba, P. Rosso, J.M. Gómez "Adapting JIRS Passage Retrieval System to the Arabic". In: Proc. 8th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2007, Springer- Verlag,LNCS(4394), pp. 530-541.

[6]Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In NAACL-HLT.

[7]. Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander Rush. Latent alignment and variational attention. In Advances in Neural Information Processing Systems, pp. 9712–9724, 2018.

[8] Burger, J. et alii. Issues, tasks, and program structures to roadmap research in question & answering (q&a), in: NIST, 2002.

[9]Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, Technical report, OpenAI.

[10]Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

[11]Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding.
CoRR, abs/1810.04805.

[12]Martin Mirakyan, Karen Hambardzumyan, and Hrant Khachatrian. 2018. Natural language inference over interaction space: ICLR 2018 reproducibility report.
CoRR, abs/1802.03198.

[13]Shuohang Wang and Jing Jiang. 2017. A compareaggregate model for matching text sequences. In 5th International Conference on Learning Representations,
ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings.
[14]Alwakid, G., Osman, T., and Hughes-Roberts, T. (2017). Challenges in sentiment analysis for Arabic social networks. Procedia Computer Science, 117:89–100.
[15]Elnagar, A., Khalifa, Y. S., and Einea, A. (2018). Hotel Arabic-reviews dataset construction for sentiment analysis applications. In Intelligent Natural Language
Processing: Trends and Applications, pages 35–52. Springer.
[16]El-Beltagy, S. R., Khalil, T., Halaby, A., and Hammad, M. (2016). Combining lexical features and a supervised learning approach for Arabic sentiment analysis. In International Conference on Intelligent Text Processing and Computational Linguistics, pages 307–319.
Springer.

[17]Alwakid, G., Osman, T., and Hughes-Roberts, T. (2017). Challenges in sentiment analysis for Arabic social networks. Procedia Computer Science, 117:89–100.
[18]Aly, M. and Atiya, A. (2013). Labr: A large scale Arabic book reviews dataset. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics
(Volume 2: Short Papers), volume 2, pages 494–498
[19]Atoum, J. O. and Nouman, M. (2019). Sentiment Analysis of Arabic Jordanian Dialect Tweets. (IJACSA) International Journal of Advanced Computer Science and Applications, 10:256–262.
[20]Azmi, A. M. and Alzanin, S. M. (2014). Aara'–a system for mining the polarity of Saudi public opinion through e-newspaper comments. Journal of Information Science,
40(3):398–410.

[21]Mohammed, A. and Kora, R. (2019). Deep learning approaches for Arabic sentiment analysis. Social Network Analysis and Mining, 9(1):52.
[22]Nabil, M., Aly, M., and Atiya, A. (2015). ASTD: Arabic Sentiment Tweets Dataset. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language
Processing, pages 2515–2519.

[23]Mozannar, H., Hajal, K. E., Maamary, E., & Hajj, H. (2019). Neural Arabic question answering. arXiv preprint arXiv:1906.05394.
[24]Antoun, W., Baly, F., & Hajj, H. (2020). Arabert: Transformer-based model for Arabic language understanding. *arXiv preprint arXiv:2003.00104*.
[25]Khouja, J. (2020). Stance prediction and claim verification: An Arabic perspective. arXiv preprint arXiv:2005.10410.
[26]Zhang, C., & Abdul-Mageed, M. (2019). BERT-Based Arabic Social Media Author Profiling. arXiv preprint arXiv:1909.04181.
[27]Hasanain, M., & Elsayed, T. (2020). bigIR at CheckThat! 2020: Multilingual BERT for ranking Arabic tweets by check-worthiness. Cappellato et al.[10].
[28]Helwe, C., Dib, G., Shamas, M., & Elbassuoni, S. (2020, December). A Semi-Supervised BERT Approach for Arabic Named Entity Recognition. In Proceedings of the Fifth Arabic Natural Language Processing Workshop (pp. 49-57).
[29]Loginova, E., Varanasi, S., & Neumann, G. (2021). Towards end-to-end multilingual question answering. Information Systems Frontiers, 23(1), 227-241.
[30]Tahsin Mayeesha, T., Md Sarwar, A., & Rahman, R. M. (2020). Deep learning based question answering system in Bengali. Journal of Information and Telecommunication, 1-34.
[31]El Bazi, I., & Laachfoubi, N. (2019). Arabic named entity recognition using deep learning approach. International Journal of Electrical & Computer Engineering (2088-8708), 9(3).
[32]Alharbi, A. and Lee, M. (2020). Combining character and word embeddings for the detection of offensive language in arabic. OSACT, 4.
[33] Djandji, M., Baly, F., antoun, w., and Hajj, H. (2020). Multi-task learning using arabert for offensive language detection. OSACT, 4.
[34] Elmadany, A., Zhang, C., Abdul-Mageed, M., and Hashemi, A. (2020). Leveraging affective bidirectional transformers for offensive language detection. OSACT, 4.
[35] Haddad, B., Orabe, Z., Al-Abood, A., and Ghneim, N. (2020). Arabic offensive language detection with attention-based deep neural networks. OSACT, 4.

[36] Hassan, S., Samih, Y., Mubarak, H., Abdelali, A., Rashed, A., and Absar Chowdhury, S. (2020). Alt submission for osact shared task on offensive language detection. OSACT, 4.

[37] Husain, F. (2020). Osact4 shared task on offensive language detection: Intensive preprocessing based approach. OSACT, 4.

[38] Keleg, A., El-Beltagy, S. R., and Khalil, M. (2020). Asu opto at osact4 - offensive language detection for Arabic text. OSACT, 4.

[39] Saeed, H. H., Calders, T., and Kamiran, F. (2020). Ocast4 shared tasks: Ensembled stacked classification for offensive and hate speech in arabic tweets. OSACT, 4.