Assessing the Linguistic Knowledge in Arabic Pre-trained Language Models Using Minimal Pairs

Wafa Alrajhi
King Saud University
wAAlrajhi@imamu.edu.sa

Hend Al-Khalifa
King Saud University
hendk@ksu.edu.sa

AbdulMalik Al-Salman
King Saud University
salman@ksu.edu.sa

Abstract

Despite the noticeable progress that we recently witnessed in Arabic pre-trained language models (PLMs), the linguistic knowledge captured by these models remains unclear. In this paper, we conducted a study to evaluate available Arabic PLMs in terms of their linguistic knowledge. BERT-based language models (LMs) are evaluated using Minimum Pairs (MP), where each pair represents a grammatical sentence and its contradictory counterpart. MPs isolate specific linguistic knowledge to test the model's sensitivity in understanding a specific linguistic phenomenon. We cover nine major Arabic phenomena from: Verbal sentences, Nominal sentences, Adjective Modification, and Idafa construction. The experiments compared the results of fifteen Arabic BERT-based PLMs. Overall, among all tested models, CAMeL-CA and GigaBERT outperformed the other PLMs by achieving the highest overall accuracy.

1 Introduction

Recently, tremendous pre-trained neural network models existed and are used effectively in different Natural language processing (NLP) tasks. This renaissance began roughly when Google launched the Transformers architecture in 2017 (Vaswani et al., 2017). Furthermore, different models are developed after the Transformers, such as Generative pre-training (GPT) (Radford et al., 2018), GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), and Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). These models have proved their strength in many NLP tasks, such as machine translation, summarization, and sentiment analysis.

In 2019, several attempts appeared to train BERT models, specifically for the Arabic language. AraBERT was one of the first Arabic models that aimed to contribute to Arabic NLP in three different tasks: Sentiment Analysis (SA), Named Entity Recognition (NER), and Question Answering (QA) (Antoun et al., 2020). Furthermore, the end of 2020 has witnessed a race where several Arabic models were published, namely: Arabic-BERT (Safaya & Yuret, 2020), GigaBERT (Lan et al., 2020), ARBERT, and MARBERT (Abdul-Mageed & Elmadany, 2020). Also, 2021 was no less intense; several versions of AraBERT models (Antoun et al., 2020), as well as ARAELECTRA (Antoun et al., 2021), and QARiB (Abdelali et al., 2021), were published. Despite these developments, the vision remains blurred in terms of how these models analyze the language in regard to various linguistic phenomena such as syntax, semantics, and grammar, which is an open area for research.

Evaluating the linguistic knowledge of PLMs has gained popularity recently. Therefore, numerous methods were developed to test the model's linguistic competence and the acquisition of different linguistic phenomena. Humans develop the ability to distinguish between grammatical and ungrammatical sentences while they are growing. Thus, studies showed that PLMs could mimic human ability, whereas, despite having no formal grammar training, the models can distinguish between grammatical and ungrammatical sentences (Warstadt et al., 2019). Many of these studies are specified for exploring the models' linguistic knowledge in the English Language (Warstadt et al., 2020; Bouraoui et al., 2020) and Chinese Language (Xiang et al., 2021). Nevertheless, to the best of our knowledge, no study has been devoted to understanding the linguistic knowledge of Arabic pre-trained language models.

PLMs, such as BERT, assign a probability/score to a sequence of words (Xiang et al., 2021). Many studies have used these scores to rank a sentence's

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 185 - 193
December 8, 2022 ©2022 Association for Computational Linguistics

correctness and evaluate the models' knowledge (Wang et al., 2019) (Shin et al., 2019). A common method to evaluate the model's linguistic knowledge is minimal pairs (MP). MP is a set of two-sentence pairs (grammatical and ungrammatical) that is used to test the model's preferences among them. Assigning a higher score for the grammatical sentence from the MP pair verifies the model's understanding of a specific phenomenon. Each pair of sentences provided by MP minimally differs by changing one word only. This change should ensure that the grammatical rule is contrasted, whereas, the grammatical and ungrammatical sentences are balanced. Example 1 illustrates a pair of MP sentences where we provided two verbal sentences; the first one is based on correct Arabic language grammar where the verb agrees with the subject in gender. The second sentence of Example 1 presents a contrast for the rule, as the verb does not agree in gender with the subject.

Furthermore, another example of Arabic language grammar is presented in Example 2. The first sentence in Example 2 (correct) provides a verb that does not agree with the subject in number, while the second sentence contrasts the rule. As we noticed from the examples, MPs are used to prompt the analysis and subsequent improvements of PLMs (Warstadt et al., 2019) (Bouraoui et al., 2020) (Xiang et al., 2021). In addition, each pair isolates a specific phenomenon, allowing the PLM to be tested separately for each linguistic phenomenon.

Example 1:

يزرع الفلاح الشجرة *(جملة صحيحة)*

yzrE AlflAH Al$jrp

The farmer (male) plants (masculine verb) the tree *(grammatical)*

تزرع الفلاح الشجرة *(جملة خاطئة)*

tzrE AlflAH Al$jrp

The farmer (male) plants (feminine verb) the tree *(ungrammatical)*

Example 2:

ذهب الولدان برحلة *(جملة صحيحة)*

*hb AlwldAn brHlp

The boys went (single) on a trip *(grammatical)*

ذهبا الولدان برحلة *(جملة خاطئة)*

*hbA AlwldAn brHlp

The boys went (dual) on a trip (ungrammatical)

In this study, we introduce a handcrafted Arabic minimal pair MPs consisting of around 3000 sentences[1]. As each MP contains both grammatical and ungrammatical sentences, the dataset is balanced and written in Modern Standard Arabic (MSA). Moreover, since the Arabic Language is extensive and complex, we limited this study to cover nine basic Arabic syntactic, semantic, and grammatical phenomena, including: verbal sentence, nominal sentence, adjective modification, and Idafa construction. Fifteen BERT-based LMs were tested using the models' sensitivity to detect the grammatical contrast. Therefore, our contributions in this paper can be listed as follows:

1- Building the first handcrafted Arabic minimal pair MPs dataset consisting of 3000 sentences.

2- Evaluating the linguistic knowledge of fifteen Arabic PLMs.

The remainder of the paper is organized as follows: the next section discusses the basic phenomena of Arabic syntax. Then, we present a description of the existing Arabic PLMs. Next, section 4 illustrates the conducted experiments, followed by their results and discussion. Finally, Section 6 concludes the paper with limitations and future work.

2 Arabic Linguistics

Arabic is a distinctive language with unique characteristics, rich morphology, and free word ordering (Habash, N.Y., 2010). The Arabic sentence is divided into two types: the verbal sentence and the nominal sentence. For each type, there are several forms that the sentence can take and remain linguistically correct. The following subsections cover a summary of these primary forms. Additionally, the relationship between nouns, case assignment, gender, and number agreement in the sentence structure are also covered. Table 1 shows acceptable and unacceptable examples of MPs for each linguistic phenomenon that we included in this study. In each example, the underlined word represents the word that we changed to contrast the grammar.

[1] https://github.com/wafa7d/AssessingArabicBERTs

Phenomenon		Accepted Example	Unaccepted Example
Verbal Sentence	1. Agreement of the verb and subject in gender	يزرع الفلاح الشجرة yzrE AlflAH Al$jrp The farmer (male) plants (masculine) the tree	تزرع الفلاح الشجرة tzrE AlflAH Al$jrp The farmer (male) plants (feminine) the tree
	2. Disagreement of the verb and subject in number	قطف الفلاحون الثمار qTf AlflAHwn AlvmAr Peasants (plural) harvested (single) fruits	قطفوا الفلاحون الثمار qTfwA AlflAHwn AlvmAr The peasants (plural) harvested (plural) the fruits
Nominal Sentence	3. Agreement of the subject and predicate in number	الطالبتان مجدتان AlTAlbtAn mjdtAn The two students are good	الطالبة مجدتان AlTAlbp mjdtAn The student are good
	4. Agreement of the subject and predicate in gender	هذا طالب نشيط h*A TAlb n$yT This (masculine) is an active student (male)	هذه طالب نشيط h*h TAlb n$yT This (masculine) is an active student (female)
Adjective Modifications	5. Rational	المهندس البارع Almhnds AlbArE The brilliant (masculine) engineer (male)	المهندس البارعة Almhnds AlbArEp The brilliant (feminine) engineer (male)
	6. Irrational	آلات جديدة |lAt jdydp New (feminine) machines (feminine)	آلات جدد ậlạt jdd New (masculine) machines (feminine)
Idafa Construction	7. Adjective agrees with head noun in case	باب حديقة كبير bạb ḥdyqẗ kbyr Large (single) garden door (single)	أبواب حديقة كبير >bwAb Hdyqp kbyr Large (single) garden doors (plural)
	8. Adjective agrees with second noun in definiteness	قراءة العلم النافع qrA'p AlElm AlnAfE Reading beneficial knowledge	قراءة علم النافع qrA'p Elm AlnAfE Reading beneficial knowledge
	9. Adjective agrees with first noun in gender	قائد الفرقة القوي qA}d Alfrqp Alqwy Strong (masculine) squad leader(male)	قائد الفرقة القوية qA}d Alfrqp Alqwyp Strong (feminine) squad leader(male)

Table 1 Minimal Pairs (MPs) for nine linguistic phenomena of Arabic Language that were covered in this paper (the translitarion is done using Buckwalter)

2.1 Verbal Sentences

Verbal sentences can be expressed in several forms, where expressing the subject may vary in each of these forms (Habash, N.Y., 2010). In this paper, we covered the following forms of verbal sentences:

- Verbal sentence with non-pronominal subject where:
  a) The verb and subject agree in gender.
  b) The verb and subject do not agree in number.

The basic form of the verbal sentence is: Verb-Subject-Object(s), where the non-pronominal subject appears after the verb. In this case, the verb and the subject should agree in the gender, but not the number, i.e., singular, dual, and plural. Consequently, the male subject requires a male verb, e.g. (He wrote – ktb – كتب), likewise if the subject is feminine, the feminine sign should be attached to the verb, e.g. (She wrote – ktbt – كتبت).

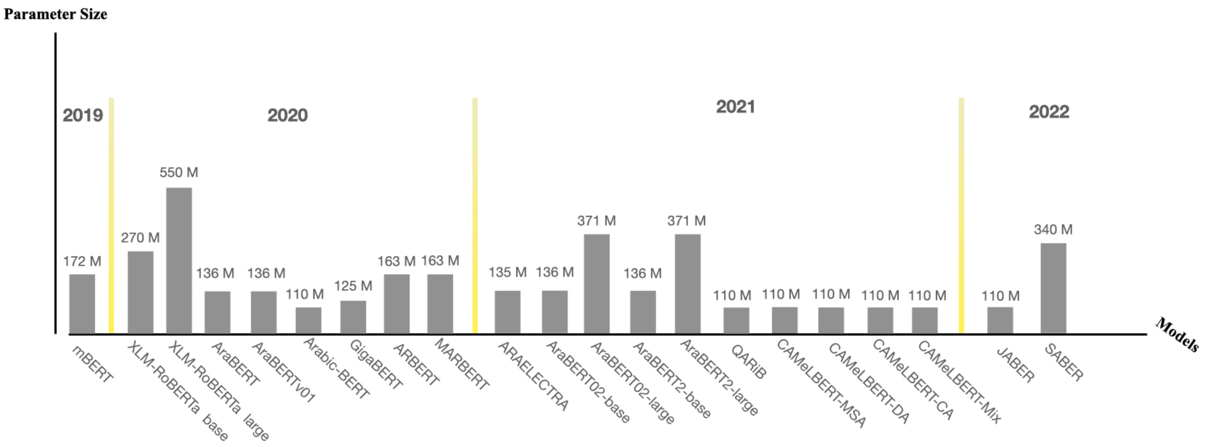

Figure 1 The Timeline of Arabic PLMs with Their Parameters

Table 1 demonstrates the different verbal sentence forms by providing acceptable and unacceptable examples for each of these forms.

2.2 Nominal Sentences

Similar to verbal sentences, nominal sentences can be expressed through different forms; the simplest form is the Subject-Predicate/Topic-Complement (Habash, N.Y., 2010). The subject can be a definite noun, proper noun, or pronoun, while the predicate is an indefinite noun, proper noun, or adjective. Two different cases are considered in the nominal sentence as follows:

1. Agreement of subject and predicate in number.
2. Agreement of subject and predicate in gender.

As mentioned above, the subject and predicate should agree in number and gender, as demonstrated by several examples in Table 1.

2.3 Adjective Modifications

Similar to English, adjectives in Arabic are nouns that describe other nouns or pronouns. The Arabic adjectives can describe rational and irrational nouns, which are the two adjective modification cases that we considered. Arabic adjectives agree in definiteness and case with nouns. However, the adjective of the rational nouns agrees in gender and number as well. Table 1 illustrates examples of the difference between rational and irrational adjectives (Habash, N.Y., 2010).

2.4 Idafa Construction

In the Idafa construction, two nouns are related; the first noun imposes the semantics and grammar on the second noun, such as: (squad leader / قائد الفرقة). It is considered a noun phrase and can be part of a second noun phrase. In this construction, an adjective might follow the Idafa construction describing the head noun, such as: (strong squad leader / قائد الفرقة القوي). This adjective agrees with the head noun in case and gender. Nevertheless, it agrees with the second noun in terms of definiteness (Habash, N.Y., 2010). The paper covers these three cases, and examples are illustrated in Table 1.

3 The Evolution of Arabic PLMs

Chronologically, the first multilingual BERT model that supported the Arabic language appeared in 2019; it was mxBERT (Pires et al., 2019). It was followed by the first monolingual Arabic model, i.e., AraBERT (Antoun et al., 2020), which appeared in early 2020. Figure 1 illustrates the evolution of Arabic PLMs, showing their parameter sizes and existence order. Moreover, Table 2 summarizes the configurations of the basic BERT-based Arabic model. Next, we provide a brief description of these PLMs.

3.1 AraBERT

AraBERT configurations followed BERT, which includes: 12 encoder blocks, 768 hidden dimensions, 12 attention heads, 512 maximum sequence lengths (Antoun et al., 2020). The masked language model task, which proves its efficiency in improving the pre-training task, was used as a pre-processing step. The total size of the pre-training dataset reached approximately 70 million sentences without any redundancy. Four empowered versions of the model were released at the beginning of 2021, where the data reached 77

Details / LM	**Variants**	**Size**	**#Words**	**#Steps**
AraBERT(Antoun et al., 2020)	MSA	23GB	2.7B	1.2M
AraBERTv01 (Antoun et al., 2020)	MSA	23GB	2.7B	1.2M
AraBERTv02-based (Antoun et al., 2019)	MSA	77GB	8.6B	3M
AraBERTv02-large (Antoun et al., 2020)	MSA	77GB	8.6B	550K
AraBERTv2-based (Antoun et al., 2020)	MSA	77GB	8.6B	550K
AraBERTv2-large (Antoun et al., 2020)	MSA	77GB	8.6B	550K
ArabicBERT (Safaya & Yuret, 2020)	MSA/DA	95GB	8.2B	4M
GigaBERT (Lan et al., 2020)	MSA	-	10.4B	1.47M
ARBERT (Abdul-Mageed & Elmadany , 2021)	MSA	61GB	6.5B	8M
MARBERT (Abdul-Mageed & Elmadany , 2021)	MSA/DA	128GB	15.6B	17M
CAMeLBERT-MSA (Inoue et al., 2021)	MSA	107GB	12.6B	1M
CAMeLBERT-DA (Inoue et al., 2021)	DA	45GB	5.8B	1M
CAMeLBERT-CA (Inoue et al., 2021)	CA	6GB	847M	1M
CAMeLBERT-Mix (Inoue et al., 2021)	MSA/ DA/CA	167GB	17.3B	1M

Table 2 BERT-based LMs Configurations

GB; AraBERT (136 million), AraBERTv01 (136 million), AraBERTv02-based (136 million), AraBERTv02-large (371 million), AraBERTv2-based (136 million), AraBERTv2-large (371 million). These versions vary in parameter size, and a more extensive dataset was used in the training process.

3.2 GigaBERT

GigaBERT is a cross-lingual model English-to-Arabic customized BERT that follows, as AraBERT, the same configuration of BERT (Antoun et al., 2020). It was trained using the fifth edition of the Gigaword English and Arabic corpora, which consists of 13 million articles. Wikipedia's data were added to manage the unbalance between English and Arabic datasets. Furthermore, the Arabic dataset was up-sampled by repeating Wikipedia's data five times and Gigaword three times.

3.3 ARBERT and MARBERT

The authors (Abdul-Mageed & Elmadany, 2021) introduced these two models, and both followed BERT architecture. ARBERT was trained on MSA only and the dataset reached 61 GB of text. MARBERT was trained on MSA and Arabic dialects, making the model more suitable for downstream tasks. Thus, almost 1 billion Arabic tweets were used to train MARBERT, which is around 128GB of text.

3.4 CAMeLBERT

The authors of (Inoue et al., 2021) proposed up to eight Arabic PLMs aiming to investigate the effect of the training data size/type variations on the behavior of these LMs. Mainly, CAMeLBERT-MSA was trained on 107 GB of Modern Standard Arabic (MSA) text, CAMeLBERT-DA was trained on 54 GB of Dialectal Arabic (DA) text, CAMeLBERT-CA was trained on 6 GB of Classical Arabic (CA) text, and CAMeLBERT-Mix is a mix of all the previous three, where its training data reached 167GB. Similar to the previous models, the authors followed the BERT model's architecture. The PLMs are evaluated on different NLP tasks: NER, POS tagging, Sentiment Analysis, dialect identification, and poetry classification. The authors elucidate the importance of the proximity of the subtask data training and pre-training data, compared to the size of the pre-training data.

4 Method

In the following subsections, we precisely describe the data coverage and the conducted experiment.

Phenomena/LMs	**AraBERT** (Antoun et al., 2020)	**AraBERTv01**(Antoun et al., 2020)	**AraBERTv02-based** (Antoun et al., 2020)	**AraBERTv02-large** (Antoun et al., 2020)	**AraBERTv2-based** (Antoun et al., 2020)	**AraBERTv2-large** (Antoun et al., 2020)	**ArabicBERT**(Safaya & Yuret, 2020)	**GigaBERT** (Lan et al., 2020)	**ARBERT** (Abdul-Mageed & Elmadany, 2021)	**MARBERT** (Abdul-Mageed & Elmadany, 2021)	**QARiB** (Abdul-Mageed et al., 2021)	**CAMeLBERT-MSA** (Inoue et al., 2021)	**CAMeLBERT-DA** (Inoue et al., 2021)	**CAMeLBERT-CA** (Inoue et al., 2021)	**CAMeLBERT-Mix** (Inoue et al., 2021)
Overall accuracy	47.7%	51.3%	52.0%	51.1%	49.7%	53.2%	52.5%	**55.1%**	53.2%	54.4%	49.2%	51.0%	49.1%	**55.1%**	49.1%
Verbal sentence															
1. Agreement of the verb and subject in gender	55.3%	44%	46%	47.3%	65.3%	58%	48.6%	60%	41.3%	39.3%	42.6%	52%	47.3%	**76.5%**	42.6%
2. Disagreement of the verb and subject in number	66.5%	60.5%	58.5%	63%	57%	55.5%	64.5%	**76%**	59.5%	62%	72%	63%	57.5%	66%	63.5%
Nominal sentence															
3. Agreement of the subject and predicate in number	**56.9%**	43%	34.4%	45%	43.7%	54.9%	54.9%	54.3%	39%	45.6%	45%	50.3%	41.7%	**56.9%**	39%
4. Agreement of the subject and predicate in gender	**56.2%**	42.7%	44.7%	41.2%	46.7%	42.7%	46.7%	46.2%	53.7%	46.2%	36.6%	44.2%	44.2%	48.2%	37.6%
Adjective Modification															
5. Rational	49.3%	51.3%	46%	45.3%	46%	50%	54%	51.3%	48%	49.3%	39.3%	47.3%	44%	**56%**	37%
6. Irrational	26%	59%	73%	63%	43%	51%	56%	64%	66%	**75%**	49%	51%	51%	56%	54%
Idafa construction															
7. Adjective agrees with head noun in case	**58%**	57%	41%	47%	52%	51%	56%	51%	48%	47%	50%	48%	39%	56%	53%
8. Adjective agrees with second noun in definiteness	49.3%	44%	53.3%	53.3%	33.3%	45.3%	45.3%	50.6%	45.3%	41.3%	46%	50.6%	**56%**	54.6%	54.6%
9. Adjective agrees with first noun in gender	56%	45.3%	42.6%	53.3%	54.6%	**58.6%**	44%	50.6%	56%	46.6%	40%	57.3%	45.3%	44%	49.3%

Table 3 Accuracy results for Arabic PLMs; **Bold** numbers indicate the highest accuracy

4.1 Data

Following Arabic basic morphology, syntax, and semantics, we constructed a handcrafted dataset for this experiment. This dataset covers nine major Arabic linguistic phenomena that include the aforementioned grammars of verbal sentences, nominal sentences, adjective modification, and Idafa. The dataset comprised well-established contrasts in Arabic Minimal Pairs (MPs), which served as a stimulus for the models, allowing us to measure the linguistic knowledge of the model. Almost 3000 MPs were constructed; 1000 MPs for the verbal structure, 1000 MPs for the nominal structure, 500 MPs for the adjective modification sentences, and 500 MPs for the Idafa construction. The data is balanced between grammatical and ungrammatical sentences, so that 50% of the data is grammatically correct. The dataset was constructed by an Arabic language expert (Master's Degree in the Arabic Language) and reviewed by three Arabic-native speakers. Accordingly, each MP belonging to the same grammar is structurally analogous, verifying that the grammatical sentence fulfills the Arabic grammar and that the ungrammatical sentence

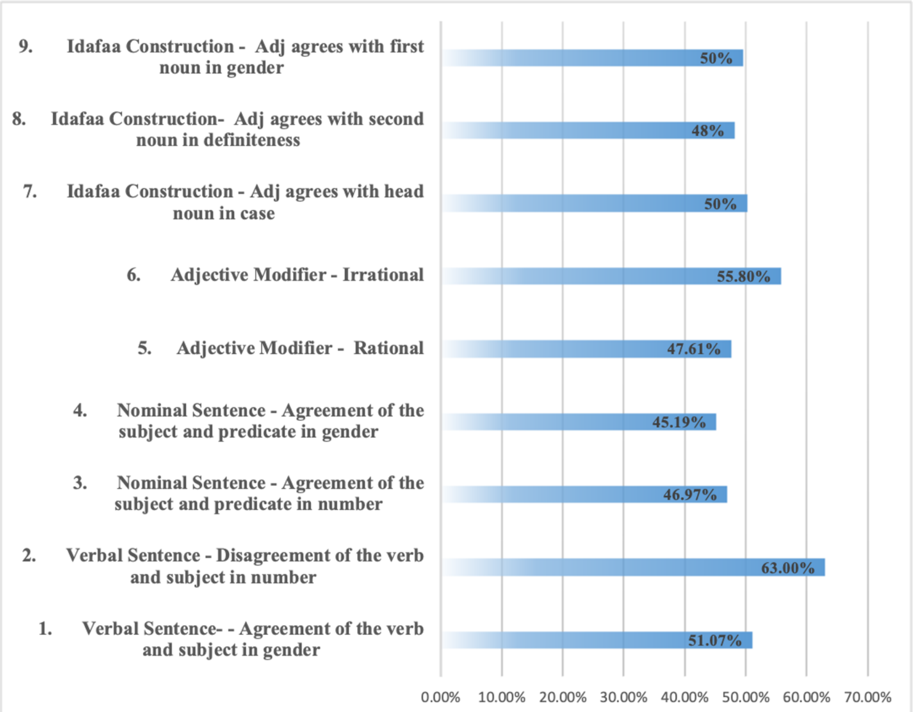

Figure 2 Arabic PLMs Average Performance on each of the Nine Arabic Phenomenon

contrasts the required grammar. All these sentences are in MSA.

4.2 Experiments

This study focuses on BERT-Based Arabic models, which allows us to examine the actual effect of different factors on the models' knowledge acquisition, such as parameter size and corpus size. As a result, we want to uncover the reasons behind the models' performance variations, if any exist. Given MPs to each model, the model should assign a higher probability to the correct grammatical sentence; in that case, the classification of MP is accepted.

In the conducted experiments, we covered fifteen Arabic PLMs. This includes six versions of AraBERT: AraBERT, AraBERTv01, AraBERTv02-base, AraBERTv02-large AraBERTv2-base, AraBERTv2-large. It also includes ArabicBERT, GigaBERT, QARiB, ARBERT, MARBERT, four versions of CAMeLBERT: CAMeLBERT-MSA, CAMeLBERT-CA, CAMeLBERTDA, and CAMeLBERT-Mix.

Table 2 illustrates the configurations of the models, highlighting the variations in terms of parameter size, corpus size, variant types of the Arabic language used in the pre-training process, and the number of training steps.

5 Results and Discussion

We evaluated each model using the accuracy metric. As shown in equation 1, the accuracy is the fraction of examples for which the model assigns a relatively higher probability for the correct sentence.

$$Accuracy = \frac{Correctly\ classified\ sentences}{Total\ number\ of\ sentences} \quad (1)$$

Table 3 illustrates the results of the Fifteen Arabic PLMs; surprisingly, even the highest accuracy did not exceed 60% in any of the covered Arabic linguistic phenomena. Overall, the performance of the models was similar, with accuracies ranging from 47% to 55%. As a result, unlike PLMs in other languages, such as English (Warstadt et al., 2020), these findings show an obvious deficiency in evaluating and understanding Arabic linguistic phenomena by PLMs.

CAMeL-CA and GigaBERT achieved the highest overall average accuracy (55.1%) in all of the Arabic phenomena we tested, including verbal, nominal, adjective, and Idafa. Unlike all models, CAMeL-CA was exclusively pre-trained on Classical Arabic, indicating that the model has acquired a better understanding of Arabic linguistic knowledge than other models. On the other hand, GigaBERT was pre-trained on Modern Standard Arabic.

For verbal sentences, CAMeL-CA yielded the highest accuracy of 76.5% in the disagreement between the subject and verb in gender, and GigaBERT outperformed all models in the agreement of verb and subject in number, with an accuracy of 76%. On the other hand, for the nominal sentence, AraBERT and CAMeL-CA performed similarly and achieved the best accuracies, approximately 57% in the subject's agreement and predicate in number. Additionally, AraBERT also achieved the highest performance in the subject's agreement and predicate in gender.

Moreover, CAMeL-CA and MARBERT have achieved the highest accuracies for the rational and irrational adjective modifiers. Specifically, CAMeL-CA achieved the highest accuracy in rational adjective modifiers, reaching 56%, While MARBERT achieved 75% accuracy in irrational. Furthermore, although the sentences in the Idafa constructions are more comprehensive, covering verbal or nominal structures, the models' accuracy remained in the same range. AraBERT, CAMeLBERT-DA, and CAMeLBERT-MSA gave the highest accuracies in the Idafa constructions.

To summarize, Figure 2 shows the average accuracy of all the models for each Arabic phenomenon. The most notable phenomenon recognized by PLMs is the disagreement between the verb and the subject in number. Conversely, the models perform poorly in the nominal sentence agreement between subject and predicate in number.

6 Conclusion

This paper aims to comprehend the linguistic abilities conferred by Arabic PLMs. We present a study to understand the basic grammar concepts obtained by the current BERT-based Arabic PLMs using MPs. Each MP represents a distinct phenomenon; hence, it can reflect the model understanding to that phenomenon. Therefore, utilizing the grammatical/ungrammatical pairs of MPs, it is feasible to assess how well the model comprehends a particular phenomenon by assigning it a higher probability to the grammatical sentence. The experiments include evaluating nine basic Arabic phenomena on fifteen BERT-based Arabic PLMs. The findings indicate a clear lack of PLMs' understanding of most of the evaluated Arabic phenomena. However, the highest average accuracy was achieved by CAMeL-CA and GigaBERT reaching 55.1%, with CAMeL-CA outperforming in three linguistic phenomena. It is worth mentioning that CAMeL-CA has used classical Arabic in its pre-training process, which justifies its high scores in our evaluation.

Finally, the capacities targeted by our experiments are not exhaustive. Future research can build on this paper's findings to study other linguistic aspects of Arabic PLMs in depth and include other models.

7 References

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Aidan N. Gomez., Lukasz Kaiser. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I., 2018. *Improving language understanding by generative pre-training*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I., 2019. *Language models are unsupervised multitask learners*. OpenAI blog, 1(8), p.9.

Brown, T., et al. 2020. *Language models are few-shot learners*. Advances in neural information processing systems, 33, pp.1877-1901.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019, June. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186).

Antoun, W., Baly, F. and Hajj, H., 2020, May. *AraBERT: Transformer-based Model for Arabic Language Understanding*. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection (pp. 9-15).

Safaya, A., Abdullatif, M. and Yuret, D., 2020. *BERT-CNN for Offensive Speech Identification in Social Media*. In Proceedings of the Fourteenth Workshop on Semantic Evaluation", Barcelona (online)", International Committee for Computational Linguistics.

Lan W, Chen Y, Xu W, Ritter A. *Gigabert: Zero-shot transfer learning from English to Arabic*. In Proceedings of The 2020 Conference on Empirical Methods on Natural Language Processing (EMNLP) 2020.

Abdul-Mageed, M. and Elmadany, A., 2021, August. *ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic*. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 7088-7105). https://doi.org/10.18653/v1/2021.acl-long.551

Antoun, W., Baly, F. and Hajj, H., 2021, April. AraELECTRA: Pre-Training Text Discriminators for Arabic Language Understanding. In Proceedings of the Sixth Arabic Natural Language Processing Workshop (pp. 191-195).

Abdelali, A., Hassan, S., Mubarak, H., Darwish, K. and Samih, Y., 2021. *Pre-training bert on arabic tweets: Practical considerations.* arXiv preprint arXiv:2102.10684..

Warstadt, A., Singh, A. and Bowman, S., 2019. *Neural Network Acceptability Judgments*. Transactions of the Association for Computational Linguistics, 7, pp.625-641. https://doi.org/10.1162/tacl_a_00290

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.F. and Bowman, S., 2020. *BLiMP: The Benchmark of Linguistic Minimal Pairs for English*. Transactions of the Association for Computational Linguistics, 8, pp.377-392. https://doi.org/10.1162/tacl_a_00321

Bouraoui, Z., Camacho-Collados, J. and Schockaert, S., 2020, April. *Inducing relational knowledge from BERT*. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 05, pp. 7456-7463). https://doi.org/10.1609/aaai.v34i05.6242

Xiang, B., Yang, C., Li, Y., Warstadt, A. and Kann, K., 2021, April. CLiMP: *A Benchmark for Chinese Language Model Evaluation*. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (pp. 2784-2790). https://doi.org/10.18653/v1/2021.eacl-main.242

Wang, A. and Cho, K., 2019, June. *BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model*. In Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation (pp. 30-36). https://doi.org/10.18653/v1/W19-2304

Shin, J., Lee, Y. and Jung, K., 2019, October. *Effective sentence scoring method using bert for speech recognition*. In Asian Conference on Machine Learning (pp. 1081-1093). PMLR.

Habash, N.Y., 2010. *Introduction to Arabic natural language processing*. Synthesis lectures on human language technologies, 3(1), pp.1-187. https://doi.org/10.2200/S00277ED1V01Y201008HLT010)

Pires, T., Schlinger, E. and Garrette, D., 2019, July. *How Multilingual is Multilingual BERT?*. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 4996-5001). https://doi.org/10.18653/v1/P19-1493

Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H. and Habash, N., 2021, April. *The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models*. In Proceedings of the Sixth Arabic Natural Language Processing Workshop (pp. 92-104).