Abstract

Recently, pretrained transformer-based architectures have proven to be very efficient at language modeling and understanding, given that they are trained on a large enough corpus. Applications in language generation for Arabic is still lagging in comparison to other NLP advances primarily due to the lack of advanced Arabic language generation models. In this paper, we develop the first advanced Arabic language generation model, AraGPT2, trained from scratch on large Arabic corpora of internet text and news articles. Our largest model, AraGPT2-MEGA, has 1.46 billion parameters, which makes it the largest Arabic language model available. We evaluate different size variants of AraGPT2 using the perplexity measure, where AraGPT2-MEGA achieves a perplexity of 29.8 on held-out articles from Wikipedia. Pretrained variants of AraGPT2 (base, medium, large, mega) are publicly available on github.com/aub-mind/arabert/aragpt2 hoping to encourage new research directions and applications for Arabic NLP.

1 Introduction

Few years ago, Natural language processing (NLP) was revolutionized with the introduction of multi-head self-attention transformer architecture (Vaswani et al., 2017). The transformer achieved superior performance compared to recurrent neural networks several NLP tasks including machine translation, sentence classification with BERT (Devlin et al., 2019), and ELECTRA (Clark et al., 2020), and sentence completion with GPT-2 (Radford et al., 2019), GROVER (Zellers et al., 2019), and CTRL (Keskar et al., 2019). Recent works have shown that larger models pre-trained on larger datasets can further improve performance i.e. RoBERTa (Liu et al., 2019), and XLM-R (Conneau et al., 2019).

GPT-1 (Radford et al., 2018) showed that Causal Language Modeling\(^1\) is an effective pretraining technique that improves a model’s generalization capabilities. GPT-2 then showed that using a larger model trained on a larger dataset surpasses the state-of-the-art of many tasks in a zero-shot setting, where a model solves a task without receiving any training on that task. Taking the scaling approach to the extreme led to the creation of GPT-3 (Brown et al., 2020), a 175 billion parameter model, also trained with CLM using terabytes of internet text. GPT-3 explored the idea of few-shot learning, where a model is given examples from a new task as a text prompt, which unlocks new capabilities at test time. It was later shown that a carefully designed GPT-3 prompt allows the model to generate website designs, scramble/uncramble words...

The advantage of scaling model sizes and training datasets comes with drawbacks, particularly the high computational cost, in addition to the huge corpora required for pre-training. It was estimated that training GPT-2 and GPT-3 costs $43K and $4.6M respectively, without any hyper-parameter tuning. These drawbacks restricted the availability of large pre-trained models to English mainly and a handful of other languages.

Work on Arabic language modeling have mostly targeted natural language understanding (NLU) by pre-training transformer-based models using the Masked Language Modeling (MLM) task i.e. ARABERT (Antoun et al., 2020). In contrast, Arabic text generation or causal language modeling haven’t received much attention. Few works such as hULMonA (ElJundi et al., 2019) used next word prediction as a pre-training task in for transfer learn-

\(^1\)This is the regular Language Modeling objective where the model learns the probability of a word given previous context. We make use of the CLM acronym to distinguish from masked language modeling (MLM).
In this paper, we develop the first advanced language generation models built from the grounds up on Arabic language. We describe the process of pretraining a GPT-2 transformer model for the Arabic language, and which we name ARAGPT2. The model comes in 4 size variants: base (135M\(^2\)), medium (370M), large (792M) and mega (1.46B\(^3\)), which allows the exploration of ARAGPT2 in multiple applications with different data availability and computational constraints. We use the perplexity measure as a way to evaluate ARAGPT2. We make the models available on github.com/aub-mind/arabert/aragpt2. Making such powerful models publicly available to the Arabic research community enables research in rising Arabic NLP fields i.e Conversational Agents (Naous et al., 2020), Detection of Automatic News Generation Detection (Harrag et al., 2020)...

Our contributions can be summarized as follows:

• A methodology to pretrain a billion-size GPT2 model on a large-scale Arabic corpus.

• Publicly releasing four variants of ARAGPT2 on popular NLP libraries.

The rest of the paper is structured as follows. Section 2 provides a concise review of previous literature on Arabic language modeling. Section 3 details the methodology used in developing ARAGPT2. Section 4 describes the experimental setup, and evaluation procedures and results. We finally conclude in Section 5.

2 Related Works

Work on Arabic causal language modeling has been mostly limited to automatic speech recognition (ASR) systems. Since the language modeling component in an ASR systems is a key module which ensures that the output text adheres with the statistical structure of language. Work on Arabic language models in ASR systems have mostly relied on N-grams language models. (Ali et al., 2014) built an N-grams language model (LM) using GALE training data transcripts of 1.4M words. More recent work in Arabic ASR implemented Recurrent neural network as a LM, using 130M tokens and achieved a perplexity of 481 compared to 436 for a 4-gram LM (Khurana et al., 2019). Hamed et al. (2017) developed a code-switched Arabic-English language model using tri-gram LM, and provided performance superior compared to two separate monolingual LMs. The code-switched LM was trained on 2.3M sentences or 13M words and achieved a perplexity of 275.

With the rising popularity of transfer learning in NLP, Arabic CLM was used as pretraining task for an Arabic universal LM, hULMonA (ElJundi et al., 2019). The model was then fine-tuned on different downstream text classification tasks. hULMonA is a 3 stack of AWD-LSTM\(^4\) layers (Howard and Ruder, 2018), trained on 600K Wikipedia article pre-segmented using the MADAMIRA Arabic morphological analyzer and disambiguator (Pasha et al., 2014).

Masked Language Modeling (MLM) has been useful as pretraining task for several Arabic NLU models. Masked Language Modeling (MLM) is a slightly different objective than CLM that requires a system to predict a masked word within a sequence compared to CLM which predict the missing word at the end of a sequence. MLM was used in models such as ARABERT (Antoun et al., 2020), Arabic-BERT (Safaya et al., 2020), Arabic-ALBERT\(^5\), GigaBERT (Lan et al., 2020), MarBERT (Abdul-Mageed et al., 2020), and QARiB (Chowdhury et al., 2020). Only two works have attempted to create a Arabic transformer causal language model. (Khooli, 2020) and (Doiron, 2020) finetuned the OpenAI GPT2-base model on Arabic Wikipedia, which was mainly trained on English text. (Doiron, 2020) also continued training on a collection of dialectal Arabic datasets, in order to create a dialectal Arabic GPT2. While this approach has shown the capability to generate Arabic text, it is sub-optimal for Arabic, and is useful in cases where the training data is scarce.

Our proposed model is hence, the first Arabic transformer-based causal language model trained from scratch on the largest Arabic corpora available at the time of writing.

\(^2\)Million Parameters
\(^3\)Billion Parameters
| Model    | Size  | Architecture | Context Size | Emb. Size | Heads | Layers | Optimizer |
|----------|-------|---------------|--------------|-----------|-------|--------|-----------|
| Base     | 135M  | GPT2          | 1024         | 768       | 12    | 12     | LAMB      |
| Medium   | 370M  | GPT2          | 1024         | 1024      | 16    | 24     | LAMB      |
| Large    | 792M  | GROVER        | 1024         | 1280      | 20    | 36     | Adafactor |
| Mega     | 1.46B | GROVER        | 1024         | 1536      | 24    | 48     | Adafactor |

Table 1: ARA-GPT2 model variants with sizes, architecture and optimizer

| Model          | Batch Size | Learning Rate | Steps | Time (days) | PPL  |
|----------------|------------|---------------|-------|-------------|------|
| Base@70K       | 1792       | 1.27e-3       | 70K   | 18H         | 59.6 |
| Base           | 1792       | 1.27e-3       | 120K  | 1.5         | 55.8 |
| Medium         | 1152       | 1e-3          | 85K   | 1.5         | 49.9 |
| Large          | 256        | 1e-4          | 220K  | 3           | 36.6 |
| Mega@450K      | 256        | 1e-4          | 450K  | 5           | 33.7 |
| Mega           | 256        | 1e-4          | 780K  | 9           | 29.8 |

Table 2: ARA-GPT2 training details and validation perplexity. Base@70K and Mega@450K models were trained for approx. half the total steps

3 ARA-GPT2: Methodology

ARA-GPT2 is a stacked transformer-decoder model trained using the causal language modeling objective. The model is trained on 77GB of Arabic text. ARA-GPT2 comes in four variants as detailed in Table 1, with the smallest model, base, having the same size as ARABERT-base which makes it accessible for the larger part of researchers. Larger model variants (medium, large, xlarge) offer improved performance but are harder to fine-tune and computationally more expensive. More details on the training procedure and dataset are provided in the following sections.

3.1 Model

ARA-GPT2 closely follows GPT2’s variant architectures and training procedure. Table 1 shows the different model sizes, number of heads, number of layers, parameter count and optimizer used for each model variant. All models are trained with context sizes of 1024 tokens. The LAMB (You et al., 2019) optimizer is used in the base and medium models only, since it allows using large batch sizes without worrying about training divergence. Using LAMB and Adam (Kingma and Ba, 2014) to train the large and mega variants isn’t possible on TPUv3 due to the optimizer’s high memory requirements, since memory cost scales linearly with the number of parameters. We overcame this limitations by following the training procedure of the GROVER model, where we make use of the Adafactor optimizer (Shazeer and Stern, 2018), which reduces memory requirements by factoring the second-order momentum parameters into a tensor product of two vectors. We also make use of the GROVER architecture instead of GPT2’s, in which the layer normalization order in the transformer block is changed.

3.2 Dataset

The training dataset is a collection of the publicly available Arabic corpora listed below:

- The unshuffled OSCAR corpus (Ortiz Suárez et al., 2020). We apply thorough filtering to the corpus to remove low quality, and short articles, in addition to removing text with inappropriate nature.
- The Arabic Wikipedia dump from September 2020.
- The 1.5B words Arabic Corpus (El-Khair, 2016)
- The OSIAN corpus (Zeroual et al., 2019).
- News articles provided by Assafir newspaper.

The total dataset size is 77GB with 8.8B words. The majority of the training data is comprised of Arabic news article, which is mostly written in MSA. The corpus also contains a small set English words i.e. named entities, which we keep without lower-casing.

---

4ASGD Weight-Dropped LSTM
5https://github.com/KUIS-AI-Lab/Arabic-ALBERT/
6word count was done after preprocessing, where a white space is inserted before and after punctuations, brackets, numbers... which increased the total word count
Preprocessing We first filter inappropriate text (short articles, articles with repeated sentences, pornographic content). Then replace links, emails and user mentions with special tokens. We remove diacritics, elongations and insert white-spaces before and after each punctuation and special characters. We also append the ‘<endoftext>’ token to the end of each article. We then train a Byte-level byte-pair-encoding (BPE) tokenizer with 64000 vocabulary size on all of our preprocessed dataset, using the optimized BPE implementation from the HuggingFace library (Wolf et al., 2020). Finally we apply BPE encoding on the preprocessed dataset, which results in a total of 9.7M training examples with 1024 sub-word tokens each.

4 Experiments and Evaluation

4.1 Pretraining Setup

All models were trained on a TPUv3-128 slice\(^7\) with different batch sizes and total number of steps as shown in Table 2. Base and mega were trained for approximately 20 epochs, while medium and large were trained for 10 and 6 epochs respectively, due to TPU access limitations.

4.2 Evaluation

We chose the perplexity score as an evaluation metric, since it measures the degree of ‘uncertainty’ a model has assigning probabilities to the test text. To make sure that the model wasn’t being tested on a text from the training set, we use Arabic Wikipedia articles that were published after August 2020, since older articles were included in the September Wikipedia dump.

Table 2 shows that, unsurprisingly, validation perplexity keeps improving with larger model sizes and longer training time. In fact, we are still underfitting the validation set from Wikipedia. We illustrate the generation capabilities of the different variants of ARAGPT2 through the selected examples in Appendix A.

4.3 Discussion

During our initial testing of the base and mega models, we noticed that the base model tends to repeat phrases more than the mega model. We solved this issue by applying a repetition penalty as in the CTRL paper (Keskar et al., 2019), which penalizes the probability scores of previously generated tokens. We also employ a ‘no repeat tri-gram’ strategy that inhibits the model from generating the same tri-gram more than once. These measures were also necessary with the mega model but only when generating very long text from short prompts.

5 Conclusion

ARAGPT2 is the first advanced Arabic language generation model based on the transformer architecture. The model was trained on the largest collection of filtered Arabic corpora. The model was evaluated using the perplexity measure which measures how well a probability model predicts a sample. Results show that ARAGPT2 is able to produce high quality Arabic text it is coherent, grammatically correct and syntactically sound.

We hope that publicly releasing ARAGPT2 will open up new research possibilities for Arabic NLP. We are also aware of the ethical implications and dangers of the malicious use of such powerful models i.e. automatic fake news generation, modeling the dataset inherent biases... Hence we are currently working on developing and releasing a detector model that is tasked to detect output generated by all ARAGPT2 variants.

Acknowledgments

This research was supported by the TFRC program, which we thank for the free access to cloud TPUs. We thank As-Safir newspaper for the data access, and also thank Nick Doiron for the insightful discussions.

References

Muhammad Abdul-Mageed, Chiyu Zhang, Abdel-Rahim Elmadany, and Lyle Ungar. 2020. Toward micro-dialect identification in diaglossic and code-switched environments. arXiv preprint arXiv:2010.04900.

Ahmed Ali, Yifan Zhang, Patrick Cardinal, Najim Da-hak, Stephan Vogel, and James Glass. 2014. A complete kaldi recipe for building arabic speech recognition systems. In 2014 IEEE spoken language technology workshop (SLT), pages 525–529. IEEE.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020, page 9.

\(^7\)TPUv3-128 has a total 2TB of HBM memory with 16GB per core. TPUs were freely provided by the TFRC program.
Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. *arXiv preprint arXiv:1804.04235*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2019. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in neural information processing systems*, pages 9054–9065.

Imad Zeroual, Dirk Goldhahn, Thomas Eckart, and Abdelhak Lakhouaja. 2019. Osian: Open source international arabic news corpus-preparation and integration into the clarin-infrastructure. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 175–182.
A  Generated Samples from ARAGPT2

Prompt

Figure 1: Random unseen context about coronavirus vaccine(top). Followed by two generated samples by ARAGPT2-mega. Generated text 1 ($top_p = 0.95$), Generated text 2 ($top_p = 1$)

Generated Text 1

Generated Text 2
يحكى أن مزارعا مخادعا قام ببيع بئر الماء الموجود في أرضه لجاره مقابل مبلغ كبير من المال، وعندما جاء المزارع الذي اشترى البارň، أذهب من هذا إلى الجوامع لكي يدعي النبي الذي لا يعنى بالماء الموجود فيه، فذهب الرجل مما ماع وتوجه إلى القاضي ليشكو المزارع المخادع. لبعد محاولات عدة لإلقائه بأن الرجل الذي يعرفه.

Generated Text 1

وأن عليه أن يغفر المبلغ الذي أشتراه به للذين استفادوا منه، وأن يشرف من المال، وأن يغفر من الماء، وأن يغفر من الرجل الذي أدفعه. فأجابه القاضي: "قد أشتراه منك البئر لكنك لم تستر بيني الماء. وفي يوم من الأيام بينما كان أحد المزارعين يتسلم على إحدى شرائح من الماء، سلّم على سبب مهنيته. فإذا الرجل. بيده يملك نقية، وافقت الرجل على شرط المزارع واعطاه الن.resolve بالماء. وقبل أن يغادر المزارع البئر قلت: إنها البئر الذي أشتراه منك، فلا تشترتب مني البئر. ففي يوم من الأيام بينما كان أحد المزارعين يجيء على إحدى شرائح من الماء، سلّم على سبب مهنيته. فإذا الرجل، بيده يملك نقية، وافقت الرجل على شرط المزارع واعطاه الن.resolve بالماء. وقبل أن يغادر المزارع البئر قلت: إنها البئر الذي أشتراه منك، فلا تشترتب مني البئر.

Generated Text 2

سماحتي يا مليكة الغابية، أرجو أن تسمعني، أنا الملك الغابى وأقوى ملكاتها، كان ذات يوم ندمت، عندما بدأ الفأر يعيش في الغابة بالضيق فذبح حوله والقفز فوقه، وإثرد أصوات مزعجة، مما أقلل من نومي ودفعه للاستيقاظ، وعندما قام من نومه كان غاضبا، فوضعت قباتته اللكى فوالفأر، وزمجر وفتل فمه ينوي ابتلاع الفأر الصغير بلقمى واحدة، صتا الفأر عندها بصوت يرتجل من الخول، راجيا أن يعفو على الفأر الصغير، وقال:

"ستامحني يا ملك الغاب، أرجوك ستامحني، لأنني لا استطيع العيش بدونك " فأجاب الملك الغابى، "أنا ملك الغابى، وأنم أيها الفأر الصغير، مجرد تابعي لي، ولا تستطيع عمل أي شيء دون إذني، أما أنا فأستطيع أن أعمل ما أشتاء وأعيش كما أشتاء، ولكن بشتر واحد هو أن تستاعدني على القتاء على جميع الفئران التي تعيش في الغابة، فقلل الفأر: "أنت ملك الغابى في الغابة، وسألته إلى هنا تعيش معهكل من طعامك، وشربك من مخرى، وأنت في الغابة ؟"، ضحك الملك الغابى على ذلك، فقال: "هذا معنى هذا الكلام من قبل، ولكن لا نأك لك ذلك كل الكلام الذي أوفره، وسواه، وأن كل سلطان على الحيوانات، كلها، أرجو أن تسمح لي بأن أعيش به سكينة وحيد وحيد الحياة سلطنتك، ونجاحك، فشك في أن أقدر به، ويجب على النبي الذي لا يعنى بالماء، أن يغفر من الماء، بل يغفر من الرجل الذي أدفعه. فأجابه القاضي: "قد أشتراه منك البئر لكنك لم تستر بيني الماء، فلا تشترتب مني البئر. ففي يوم من الأيام بينما كان أحد المزارعين يجيء على إحدى شرائح من الماء، سلّم على سبب مهنيته. فإذا الرجل، بيده يملك نقية، وافقت الرجل على شرط المزارع واعطاه الن.resolve بالماء. وقبل أن يغادر المزارع البئر قلت: إنها البئر الذي أشتراه منك، فلا تشترتب مني البئر.

Figure 2: Random unseen contexts about children stories. Followed by a generated sample by ARA GPT2-mega with top_p = 0.95.
فتة الدجاج اللبنانية

الطريقة التحضير
نضع الدجاج في قدر من الماء مع البصل والهيل وورق الغار والملح واللفل الأسود ونسلقه على نار متوسطة حتى ينصب تماما. صفي الدجاج من المرق ونضعه جانبا ثم صفي المرق ونضعه في وعاء. نحمي الزيت ونizziness فيه الخبز.

توجه شرائح الفليفلة الخضراء والخضرة والخضرة والبندورة المقشرة والمفرومة فوق الخبز، ثم نرش عليها البقدونس المفروم والصوص الحلو. ندهن صلصة فرن بالقليل من الزيت ونرش عليها الزيت من الملحين ثم ندخل الصينية في فرن محمي مسبقًا على حرارة 180 درجة مئوية لمدة تتراوح بين 15 و 20 دقيقة أو حتى يحمر وجه الصينية. نخرج الصينية من الفرن ونتركها لتردد قليلا ثم نقطعها و نقدمها ساخنة.

مرحبا كيف حالك؟

ان شاء الله بخير. اليوم جبتلكم لعبة حلوة كثير و هي عبارة عن مجموعة من الكرات الملونة بألوان مختلفة، كل ما عليك فعله هو الضغط على الكرة التي تريد أن ترميها في المكان المخصص لها ثم أضغط عليها مرة أخرى

لبنان تعد الجمهورية اللبنانية دولة عربية عاصمتها بيروت، وتقع على الساحل الشمالي للبحر الأطلسي المتوسط، وتعد واحدة من أكثر البلدين المكتنزة بالسكان، وأنه إلى تاريخها فهي تعتبر موطناً لأقدم الحضارات البشرية في الألفية الثالثة قبل الميلاد. تعتبر المنطقة الساحلية جزءًا من المناطق البلدية، وتشتهر بالكثير من العيوقة والجمهورية، والتي دفعت الناس إلى الاعتراف بها كواحدة من أهم المراكز الثقافية والتجارية في ذلك الوقت، وكان تضاريسها وعاء دوارة، حيث أنها تمثلت في تاريخ في السنوات، وأعلنت كدولة جمهورية عام 1926 م، وفي عام 1946 حصلت لبنان على استقلاله بعد الاستعمار الفرنسي.

[[ Леbnan

2020]]

[1] تحت اللغة العربية هي اللغة الرسمية للبلد، بالإضافة إلى وجود تجنيز رسميين أخرين هما الإنجليزية والفرنسية. أما اللغة السائدة في البلاد فإنها الإنجليزية، حيث يتكلم المسلمين حوالي 96% من اجتماعي العاطف، ويتوزع الباقون على الطبقات الصحية والدينية الأخرى، ومن أبرز المصالح السياحية في لبنان مدينة بيروت التي تعد عاصمة البلاد مركزاً للثقافة والاقتصاد، مما يعد بها العديد من المواقع الأثرية الهامة مثل مغارة جعيتا، وقنعة بعلبك، وبرج إيليل، وتحتراج سرق، وقصر بنين الذي يعد واحدًا من أهم المعالم الأثرية في لبنان.

[2] تحت الدينية في لبنان بشكل أساسي على الطيور الضيافة والمناخ المعتدل، وتعتبر الطيور الممتعة في المناطق الجبلية والساحلية للاستمتاع بالمناظر الطبيعية الجبلية والمطهرة، بينما يفضل البعض الآخر التوجه إلى الوديان والمناطق الداخلية للتمتع بمناخها المعتدل والاستمتاع في أحلى الطيور الممتعة.