AraBART: a Pretrained Arabic Sequence-to-Sequence Model for Abstractive Summarization

Anonymous ACL submission

Abstract

Like most natural language understanding and generation tasks, state-of-the-art models for summarization are transformer-based sequence-to-sequence architectures that are pretrained on large corpora. While most existing models focused on English, Arabic remained understudied. In this paper we propose AraBART, the first Arabic model in which the encoder and the decoder are pretrained end-to-end, based on BART (Lewis et al., 2020). We show that AraBART achieves the best performance on multiple abstractive summarization datasets, outperforming strong baselines including a pretrained Arabic BERT-based model and multilingual mBART and mT5 models.

1 Introduction

Summarization is the task of transforming a text into a shorter representation of its essential meaning in natural language. Extractive approaches (Nallapati et al., 2017; Narayan et al., 2018b; Zhou et al., 2018; See et al., 2017) identify informative spans in the original text and stitch them together to generate the summary. Abstractive approaches on the other hand are not restricted to the input (Rush et al., 2015; Chopra et al., 2016; Dou et al., 2021).

While the vast majority of published models in both categories focused on English, some tackled other languages including Chinese (Hu et al., 2015) and French (Kamal Eddine et al., 2021b), while Arabic remained understudied. In fact, most Arabic summarization models are extractive (Qassem et al., 2019; Alshanqiti et al., 2021). They generate explainable and factual summaries but tend to be verbose and lack fluency. Addressing this problem, abstractive models are flexible in their word choices, resorting to paraphrasing and generalization to obtain more fluent and coherent summaries. Sequence-to-sequence (seq2seq) is the architecture of choice for abstractive models. Al-Maleh and Desouki (2020), for instance, apply the pointer-generator network (See et al., 2017) to Arabic, while Khalil et al. (2022) propose a more generic RNN-based model.

There are, however, two main issues with abstractive models as applied to Arabic. First, they are trained and evaluated either on extractive datasets such as KALIMAT (El-Haj and Koulali, 2013) and ANT Corpus (Chouigui et al., 2021), or on headline generation datasets such as AHS (Al-Maleh and Desouki, 2020), which only contains short and rather extractive headlines. Second, despite their state-of-the-art performance, abstractive models frequently generate content that is non-factual or unfaithful to the original text. Maynez et al. (2020) showed that English models that are based on the Transformer architecture such as BERT2BERT (Rothe et al., 2020) efficiently mitigate this phenomenon thanks to pretraining on large corpora. Therefore, Elmadani et al. (2020) finetuned a pretrained BERT using the encoder-decoder architecture of BERTSUM (Liu and Lapata, 2019). However, only the encoder is pretrained, the decoder and the connection weights between the encoder and the decoder are initialized randomly which is sub-optimal.

To address these two problems, we propose AraBART, the first sequence-to-sequence Arabic model in which the encoder, the decoder and their connection weights are pretrained end-to-end using BART's denoising autoencoder objective (Lewis et al., 2020). While the encoder is bidirectional, the decoder is auto-regressive and thus more suitable for summarization than BERT-based decoders. We finetuned and evaluate our model on two abstractive datasets. The first is Arabic Gigaword (Parker et al., 2011), a newswire headline-generation dataset, not previously exploited in Arabic abstractive summarization; the second is XL-Sum, a multilingual text summarization dataset for 44 languages including Arabic (Hasan et al., 2021). AraBART achieves state-of-the-art results outperforming pretrained BERT-based models as well as a much larger model, mBART25 (Liu et al., 2020), a multilingual denoising auto-encoder pretrained on 25 different languages using the BART objective.

In section 2 we present the architecture and the pretraining settings of AraBART. In section 3 we evaluate and compare AraBART against three strong baselines on a wide range of abstractive summarization datasets. Finally, we discuss related work in section 4.

2 AraBART

AraBART follows the architecture of BART Base (Lewis et al., 2020), which has 6 encoder and 6 decoder layers and 768 hidden dimensions. In total AraBART has 139M parameters. We add one additional layer-normalization layer on top of the encoder and the decoder to stabilize training at FP16 precision, following (Liu et al., 2020). We use sentencepiece (Kudo and Richardson, 2018) to create the vocabulary of AraBART. We train the sentencepiece model on a randomly sampled subset of the pretraining corpus (without any preprocessing) with size 20GB. We fix the vocabulary size to 50K tokens and the character coverage to 99.99% to avoid a high rate of unknown tokens.

2.1 Pretraining

We adopt the same corpus used to pretrain AraBERT (Antoun et al., 2020). While Antoun et al. (2020) use a preprocessed version of the corpus, we opted to reverse the preprocessing by using a script that removes added spaces around non alphabetical characters, and also undo some words segmentation. The use of a corpus with no preprocessing, makes the text generation more natural. The size of the pretraining corpus before/after sentencepiece tokenization is 73/96 GB.

Pretraining Objective AraBART is a denoising autoencoder i.e. it learns to reconstruct a corrupted text. The noise function applied to the input text are the same as in Lewis et al. (2020). The first noise function is *text infilling*, where 30% of the text is masked by replacing a number of text spans with a [MASK] token. The length of the spans is sampled from a Poisson distribution with $\lambda = 3.5$. The second noise function is *sentence permutation*, where the sentences of the input text are shuffled based on the full stops.

Pretraining Settings AraBART pretraining took approximately 60h. The pretraining was carried out on 128 Nvidia V100 GPUs which allowed for 25 full passes over the pretraining corpus. We used the Adam optimizer with $\epsilon = 10^{-6}$, $\beta_1 = 0.9$, and $\beta_2 = 0.98$ following Liu et al. (2019). We use a warm up for 6% of the pretraining were the learning rate linearly increases from 0 to 0.0006, then decreases linearly to reach 0 at the end of the pretraining. We fixed the update frequency to 2 and we use a dropout 0.1 in the first 20 epochs and we changed it to 0 in the last 5 epochs. Finally we used FP16 to speed-up the pretraining. The pretraining is done using Fairseq (Ott et al., 2019).

3 Experiments

3.1 Datasets

To evaluate our model, we use several datasets that consist mostly of news articles annotated with summaries with different level of abstractivness. The first 7 datasets (*AAW*, *AFP*, *AHR*, *HYT*, *NHR*, *QDS* and *XIN*) are subsets of the Arabic Gigaword (Parker et al., 2011) corpus. Each one is a different news source, composed of document-headline pairs. In all these datasets we use a train set of 50K examples, a validation set of size 5K examples and a test set of size 5K examples, selected randomly. The *MIX* dataset consists of 60K examples uniformly sampled from the union of the 7 different sources.

In addition the Arabic Gigaword corpus, we use XL-Sum (Hasan et al., 2021). The news articles in XL-sum are annotated with summaries and titles, thus creating two tasks: summary and title generation.

Table 1 shows that the different datasets used in our experiments cover a wide range of article/summary lengths and levels of abstractivness.

3.2 Baselines

We compare our model to three types of state-of-the-art baselines. The first, called C2C, is a monolingual seq2seq model based on BERT2BERT (Rothe et al., 2020). The encoder and decoder are initialized using CAMELBERT (Inoue et al., 2021) weights while the cross-attention weights are randomly initialized.[1] C2C has 246M parameters in total.

The second baseline is mBART25 (Liu et al., 2020) which is a multilingual BART pretrained on

[1] We experimented with ARABERT (Antoun et al., 2020) which was slower to converge and didn't achieve better performance.

		Datasets									
		AAW	*AHR*	*AFP*	*HYT*	*NHR*	*QDS*	*XIN*	*MIX*	*XL-S*	*XL-T*
Average #tokens	*document*	453.3	394.2	232.8	474.0	455.9	450.6	187.2	364.5	428.7	428.7
	summary	15.5	9.2	8.3	11.2	10.4	8.0	8.2	9.4	25.6	9.4
%novel n-grams in summary	*unigrams*	44.2	46.5	30.7	42.4	46.5	24.9	26.4	40.0	53.5	44.3
	bigrams	78.5	78.4	63.6	78.6	80.7	46.9	48.5	72.2	85.8	81.2
	trigrams	91.2	91.3	81.9	92.0	92.8	57.5	60.8	86.3	95.2	94.1

Table 1: Statistics of Gigaword subsets an XL-Sum summaries (XL-S) and titles (XL-T). The first two lines show the average document and summary lengths. The percentage of n-grams in the summary that do not occur in the input article is used as a measure of abstractiveness (Narayan et al., 2018a).

25 different languages including Arabic. Although mBART25 was initially pretrained for neural machine translation, it was shown that it can be used in monolingual generative tasks such as abstractive summarization (Kamal Eddine et al., 2021b). mBART25 has 610M parameters in total.

While mBART25 is pretrained on multilingual corpora, we finetuned it on Arabic data only. We therefore, include a third multilingual baseline pretrained and finetuned on multilingual data. We use the checkpoint[2] of mT5$_{base}$ in the comparison on XL-S (summary). This checkpoint was finetuned on the training set of the 45 different languages included in the corpus. The total training size is 1M multilingual examples shuffled together (Hasan et al., 2021). mT5$_{base}$ has 582M parameters in total.

3.3 Training and Evaluation

We finetuned each model for three epochs, using the Adam optimizer and 5×10^{-5} maximum learning rate with linear decay scheduling. In the generation phase we use beam-search with beam size of 3.

For evaluation, we first normalize the output summaries as is standard practice in Arabic: we removed Tatweel and diacritization, we normalized Alef/Yaa and separated punctuations. We report ROUGE-1, ROUGE-2 and ROUGE-L f1-scores (Lin, 2004). However, these metrics are solely based on surface-form matching and have limited sense of semantic similarity (Kamal Eddine et al., 2021a). Thus we opted for using BERTScore (Zhang et al., 2020), a metric based on the similarity of the contextual embeddings of the reference and candidate summaries, produced by a BERT-like model.[3]

3.4 Results

We observe in Table 2 that AraBART outperforms C2C on all datasets with a clear margin. This is probably a direct consequence of pretraining the seq2seq architecture end-to-end.

AraBART also outperforms mBART25 on XL-Sum which is the most abstractive dataset. On Gigawords, AraBART is best everywhere except on AHR with mitigated results. On QDS, however, it falls clearly behind mBART25 on all metrics. In fact, we notice that the gap between AraBART and the baselines is greater on the XL-Sum dataset than Gigaword. For instance, our model's ROUGE-L score is 2.9 absolute points higher that mBART25 on XL-S while the maximum margin obtained on a Gigaword subset is 1.4 points on AAW and HYT. We observe a tendency for AraBART to outperform mBART on more abstractive datasets. In fact, the margin between their BERTScores is positively correlated with abstractiveness as measures by the percentage of novel trigrams.[4]

On the XL-Sum dataset, AraBART also outperforms mT5 which was finetuned in multilingual setup using more data (Hasan et al., 2021).

Figure 1 presents some examples of the output of the various systems we studied.

4 Related Work

Arabic Summarization The overwhelming majority of past Arabic models are extractive (Douzidia and Lapalme, 2004; Azmi and Al-thanyyan, 2009; El-Haj et al., 2011; El-Shishtawy and El-Ghannam, 2012; Haboush et al., 2012; Belkebir and Guessoum, 2015; Qaroush et al.,

[2]https://huggingface.co/csebuetnlp/mT5_multilingual_XLSum

[3]We use the official implementation (https://github.com/Tiiiger/bert_score) with the following options: -m UBC-NLP/ARBERT -l 9 (Chiang et al., 2020)

[4]With a Pearson R score of 0.6625 and p-value<0.05.

Source	Model	R1	R2	RL	BS
AAW	AraBART	**30.7**	**15.3**	**27.4**	**62.5**
	mBART25	29.5	14.4	26.0	61.5
	C2C	24.6	9.87	21.7	58.3
AFP	AraBART	**55.0**	**37.9**	**53.4**	**77.5**
	mBART25	54.8	37.3	52.8	77.2
	C2C	50.0	32.2	48.4	74.8
AHR	AraBART	**39.1**	25.4	**37.7**	**68.2**
	mBART25	**39.1**	**26.1**	37.5	68.1
	C2C	33.0	19.7	31.8	63.5
HYT	AraBART	**33.1**	**17.5**	**30.7**	**63.8**
	mBART25	32.0	16.2	29.3	63.1
	C2C	27.4	11.5	25.2	59.6
NHR	AraBART	**32.0**	**17.2**	**30.3**	**61.2**
	mBART25	31.0	16.2	29.2	60.3
	C2C	24.1	10.0	22.9	53.0
QDS	AraBART	62.1	53.9	61.4	80.3
	mBART25	**62.4**	**54.1**	**61.7**	**80.4**
	C2C	57.9	48.9	57.4	77.3
XIN	AraBART	**66.0**	**53.9**	**65.1**	**84.4**
	mBART25	65.1	53.4	64.2	84.0
	C2C	62.4	50.1	61.6	82.5
MIX	AraBART	**39.2**	25.5	**37.6**	**67.6**
	mBART25	39.0	**25.6**	37.1	67.2
	C2C	32.8	19.1	31.4	62.5
XL-S	AraBART	**34.5**	**14.6**	**30.5**	**67.0**
	mBART25	32.1	12.5	27.6	65.3
	C2C	26.9	8.7	23.1	61.6
	mT5_{base}	32.8	12.7	28.7	65.8
XL-T	AraBART	**32.0**	**13.7**	**29.4**	**65.8**
	mBART25	29.8	11.7	26.9	64.3
	C2C	25.2	7.9	22.9	61.1
Macro Averages	AraBART	**42.4**	**28.8**	**40.3**	**69.8**
	mBART25	41.5	28.1	39.2	69.1
	C2C	36.4	23.1	34.6	65.4

Table 2: The performance of AraBART, mBART25 and C2C (CamelBert2CamelBert) on all datasets in terms of ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL) and BERTScore (BS). Macro averages are computed over all datasets.

2021; Ayed et al., 2021). Recently, seq2seq abstractive models for Arabic have been proposed in the literature (Al-Maleh and Desouki, 2020; Suleiman and Awajan, 2020; Khalil et al., 2022), but none of them used pretraining. Fine-tuning Transformer-based language models like BERT (Devlin et al., 2019) has been shown to help Arabic abstractive (Elmadani et al., 2020) and extractive (Helmy et al., 2018) summarization, but unlike AraBART, not all components of the model are pretrained. Readily-available multilingual pretrained seq2seq models have been applied to Arabic summarization. Kahla et al. (2021) uses mBART25 (Liu et al., 2020) in cross-lingual transfer setup on an unpublished dataset, while Hasan et al. (2021) experiment with mT5 (Xue et al., 2021) on XL-Sum. Our model, tailored specifically for Arabic, outperform mBART25 and mT5 for almost all datasets despite having a smaller architecture with less parameters.

Arabic Datasets Most available datasets for Arabic are extractive (El-Haj et al., 2010; Chouigui et al., 2021), use short headlines that are designed to attract the reader (Webz.io; Al-Maleh and Desouki, 2020), or contain machine-generated (El-Haj and Koulali, 2013) or translated (El-Haj et al., 2011) summaries. Notable exceptions we choose for our experiments are Gigaword (Parker et al., 2011) and XL-Sum (Hasan et al., 2021) because they cover both headline and summary generation, contains multiple sources, and manifest variable levels of abstractivness as shown in Table 1.

Pretrained seq2seq models BART-based models have been developed for multiple language including English (Lewis et al., 2020), French (Kamal Eddine et al., 2021b) and Chinese (Shao et al., 2021) in addition to multilingual models (Liu et al., 2020). While they can be finetuned to perform any language understanding or generation tasks, we focus on summarization in this work.

5 Conclusion and Future Work

We release AraBART, the first sequenece-to-sequence pretrained Arabic model. We evaluated our model on a set of abstractive summarization tasks, with different level of abstractiveness. We compared AraBART to two state-of-the-art models and we showed that it outperforms them almost everywhere despite the fact that it is smaller in terms of parameters. In future work, we are planning to extend the model to multitask setups to take advantage of availability of both titles and summaries in some datasets including XL-Sum, and use external knowledge sources to improve faithfulness. We will also explore new directions for evaluating summarization on morphologically rich languages like Arabic.

Ethical Considerations

Limitations Our models are optimized for news text summarization; we do not expect compararble performance on other summarization tasks without additional training data.

Risks We acknowledge that our models sometimes produce incorrect non-factual and non-grammatical output, which can be misleading to general users.

Data All of the data we used comes from reputable news agencies and do not contain unanonymized private information or malicious social media content.

Models We will make our pretrained and fine-tuned models available on the well known Hugging Face models hub[5], so it can be easily used and distributed for research or production purposes.

References

Molham Al-Maleh and Said Desouki. 2020. Arabic text summarization using deep learning approach. *Journal of Big Data*, 7:1–17.

Abdullah Alshanqiti, Abdallah Namoun, Aeshah Alsughayyir, Aisha Mousa Mashraqi, Abdul Rehman Gilal, and Sami Saad Albouq. 2021. Leveraging distilbert for summarizing arabic text: An extractive dual-stage approach. *IEEE Access*, 9:135594–135607.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. 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*, pages 9–15, Marseille, France. European Language Resource Association.

Alaidine Ben Ayed, Ismaïl Biskri, and Jean-Guy Meunier. 2021. Arabic text summarization via knapsack balancing of effective retention. *Procedia Computer Science*, 189:312–319. AI in Computational Linguistics.

Aqil Azmi and Suha Al-thanyyan. 2009. Ikhtasir — a user selected compression ratio arabic text summarization system. In *2009 International Conference on Natural Language Processing and Knowledge Engineering*, pages 1–7.

Riadh Belkebir and Ahmed Guessoum. 2015. A supervised approach to arabic text summarization using adaboost. In *New Contributions in Information Systems and Technologies*, pages 227–236, Cham. Springer International Publishing.

[5]https://huggingface.co/models

Cheng-Han Chiang, Sung-Feng Huang, and Hung-yi Lee. 2020. Pretrained language model embryology: The birth of ALBERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6813–6828, Online. Association for Computational Linguistics.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California. Association for Computational Linguistics.

Amina Chouigui, Oussama Ben Khiroun, and Bilel Elayeb. 2021. An arabic multi-source news corpus: Experimenting on single-document extractive summarization. *Arabian Journal for Science and Engineering*, 46(4):3925–3938.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. GSum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842, Online. Association for Computational Linguistics.

Fouad Douzidia and Guy Lapalme. 2004. Lakhas, an arabic summarization system. In *Proceedings of DUC'04*, pages 128–135, Boston. NIST, NIST.

M. El-Haj, Udo Kruschwitz, and C. Fox. 2010. Using mechanical turk to create a corpus of arabic summaries. In *Proceedings of the 7th International Conference on Language Resources and Evaluation : Workshops & Tutorials May 17-18, May 22-23, Main Conference May 19-21, Valletta*. ELRA, Paris.

Mahmoud El-Haj and Rim Koulali. 2013. Kalimat a multipurpose arabic corpus. pages 22–25. Second Workshop on Arabic Corpus Linguistics (WACL-2).

Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. 2011. Exploring clustering for multi-document arabic summarisation. In *Information Retrieval Technology - 7th Asia Information Retrieval Societies Conference, AIRS 2011, Dubai, United Arab Emirates, December 18-20, 2011. Proceedings*, volume 7097 of *Lecture Notes in Computer Science*, pages 550–561. Springer.

Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. 2011. Multi-document arabic text summarisation. *2011 3rd Computer Science and Electronic Engineering Conference (CEEC)*, pages 40–44.

Tarek El-Shishtawy and Fatma El-Ghannam. 2012. Keyphrase based arabic summarizer (kpas). In *2012 8th International Conference on Informatics and Systems (INFOS)*, pages NLP–7–NLP–14.

Khalid N. Elmadani, Mukhtar Elgezouli, and Anas Showk. 2020. BERT fine-tuning for arabic text summarization. *CoRR*, abs/2004.14135.

Ahmad Haboush, Ahmed Momani, Maryam Al-Zoubi, and Motassem Al-Tarazi. 2012. Arabic text summerization model using clustering techniques. *World Comput Sci Inf Technol J*, 2.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Muhammad Helmy, R. M. Vigneshram, Giuseppe Serra, and Carlo Tasso. 2018. Applying deep learning for arabic keyphrase extraction. In *Fourth International Conference On Arabic Computational Linguistics, ACLING 2018, November 17-19, 2018, Dubai, United Arab Emirates*, volume 142 of *Procedia Computer Science*, pages 254–261. Elsevier.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale Chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1972, Lisbon, Portugal. Association for Computational Linguistics.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Mram Kahla, Zijian Győző Yang, and Attila Novák. 2021. Cross-lingual fine-tuning for abstractive Arabic text summarization. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 655–663, Held Online. INCOMA Ltd.

Moussa Kamal Eddine, Guokan Shang, Antoine J-P Tixier, and Michalis Vazirgiannis. 2021a. Frugalscore: Learning cheaper, lighter and faster evaluation metricsfor automatic text generation. *arXiv preprint arXiv:2110.08559*.

Moussa Kamal Eddine, Antoine Tixier, and Michalis Vazirgiannis. 2021b. BARThez: a skilled pretrained French sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9369–9390, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ahmed Mostafa Khalil, Y. M. Wazery, Marwa E. Saleh, Abdullah Alharbi, and Abdelmgeid A. Ali. 2022. Abstractive arabic text summarization based on deep learning. *Computational Intelligence and Neuroscience*, 2022:1566890.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3075–3081. AAAI Press.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018a. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018b. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Robert Parker, David Graff, Ke Chen, Junbo Kong, and Kazuaki Maeda. 2011. Arabic gigaword fifth edition. https://doi.org/10.35111/p02g-rw14.

Aziz Qaroush, Ibrahim Abu Farha, Wasel T. Ghanem, Mahdi Washaha, and Eman Maali. 2021. An efficient single document arabic text summarization using a combination of statistical and semantic features. *J. King Saud Univ. Comput. Inf. Sci.*, 33:677–692.

Lamees Al Qassem, Di Wang, Hassan Barada, Ahmad Al-Rubaie, and Nawaf Almoosa. 2019. Automatic Arabic text summarization based on fuzzy logic. In *Proceedings of the 3rd International Conference on Natural Language and Speech Processing*, pages 42–48, Trento, Italy. Association for Computational Linguistics.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation.

Dima Suleiman and Arafat Awajan. 2020. Deep learning based abstractive arabic text summarization using two layers encoder and one layer decoder. *Journal of Theoretical and Applied Information Technology*, 98:3233.

Webz.io.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663, Melbourne, Australia. Association for Computational Linguistics.

A Example Appendix

Figure 1 presents some examples of the output of the various systems we studied.

		Arabic	English
(a)	**Reference**	تنظر محكمة عسكرية امريكية في وقت لاحق من اليوم في قضية الطبيب النفسي العسكري الامريكي ، نضال حسن ، الذي اعترف بقتل 13 شخصا واصابة اكثر من ثلاثين اخرين في اطلاق نار بقاعدة فورت هود منذ اربعة اعوام .	A US military court will hear, later in the day, the case of the US military psychiatrist, Nidal Hassan, who confessed to killing 13 people and wounding more than thirty others in a shooting at Fort Hood base four years ago.
	AraBART	مثل نضال حسن ، المتهم بقتل 13 جنديا امريكيا في قاعدة عسكرية في افغانستان ، امام محكمة عسكرية في مدينة فروت هود .	Nidal Hassan, accused of killing 13 US soldiers at a military base in Afghanistan, appeared before a military court in Fort Hood.
	mBART25	اعترف رجل افغاني بقتل 13 جنديا امريكيا وجرح اخرين في قاعدة عسكرية امريكية في ولاية فرجينيا في نوفمبر / تشرين الثاني الماضي .	An Afghan man confessed to killing 13 US soldiers and wounding others at a US military base in Virginia last November.
	C2C	قال رجل الدين الامريكي ، حسن حسن ، انه " لا يستطيع ان يقتل احد جنوده " ، بعد ان قتل جنديا امريكيا في قاعدة عسكرية	The American cleric, Hassan Hassan, said that he "could not kill one of his soldiers", after he killed an American soldier in a military base
	mT5 *base*	قضت محكمة عسكرية في ولاية كاليفورنيا الامريكية باعدام رجل قتل 13 جنديا امريكيا في قاعدة للجيش الامريكي عام 2009 .	A California military court has sentenced to death a man who killed 13 US soldiers at a US military base in 2009.
(b)	**Reference**	توفي يوم الثلاثاء عن عمر ناهز 89 عاما النجم البريطاني السير روجر مور .	British star Sir Roger Moore died on Tuesday at the age of 89.
	AraBART	توفي الممثل البريطاني روجر مور عن عمر يناهز 91 عاما بعد صراع مع مرض السرطان .	British actor Roger Moore has died at the age of 91, after a battle with cancer.
	mBART25	قالت اسرة النجم السينمائي الشهير روجر مور ان والدهم توفي عن عمر يناهز 91 عاما بعد صراع طويل مع مرض السرطان .	The family of famous film star Roger Moore said that their father passed away at the age of 91 after a long battle with cancer.
	C2C	توفي الممثل البريطاني الشهير روجر مور ، الذي اشتهر بدور جيمس بوند في فيلم " غراند سلام " ، عن عمر يناهز 91 عاما	Renowned British actor Roger Moore, best known for his role as James Bond in the movie "Grand Slam", has died at the age of 91.
	mT5 *base*	توفي السير روجر مور ، بطل دور الجاسوس جيمس بوند ، عن عمر يناهز 67 عاما في سويسرا بعد صراع قصير مع مرض السرطان .	Sir Roger Moore, star of James Bond, has died at the age of 67 in Switzerland after a short battle with cancer.
(c)	**Reference**	كشفت دراسة حديثة - تناولت الخلايا الجذعية - عن تاثير السجائر علي الاجنة .	A recent study - dealing with stem cells - revealed the effect of cigarettes on fetuses.
	AraBART	توصلت دراسة علمية الي ان تدخين الامهات اثناء الحمل له اثار ضارة علي الاجنة .	A scientific study concluded that smoking by mothers during pregnancy has harmful effects on fetuses.
	mBART25	خلصت دراسة علمية حديثة الي ان تدخين الامهات السجائر له اثار ضارة علي اجنة الذكور واجنة الاناث .	A recent scientific study concluded that cigarette smoking by mothers has harmful effects on male fetuses and female fetuses.
	C2C	قال علماء ان تدخين السجائر في الامعاء قد يكون له تاثير علي صحة الانسان .	Scientists said that smoking cigarettes in the intestines may have an effect on human health.
	mT5 *base*	اظهرت دراسة حديثة ان السجائر قد يؤدي تدخين الامهات الي اضرار كبيرة علي الاجنة .	A recent study showed that smoking by mothers may cause significant harm to fetuses.

Figure 1: Three selected examples contrasting the output of the various systems we studied. All examples are from the XL-Sum summaries test set. We provide English translations to provide context for the general readers.