ARABIC SOFT SPELLING CORRECTION WITH T5

Mohammed Al-Qaraghuli and Ola Arif Jaafar

(Received: 12-Nov.-2023, Revised: 3-Jan.-2024, Accepted: 20-Jan.-2024)

ABSTRACT

Spelling correction is considered a challenging task for resource-scarce languages. The Arabic language is one of these resource-scarce languages, which suffers from the absence of a large spelling correction dataset, thus datasets injected with artificial errors are used to overcome this problem. In this paper, we trained the Text-to-Text Transfer Transformer (T5) model using artificial errors to correct Arabic soft spelling mistakes. Our T5 model can correct 97.8% of the artificial errors that were injected into the test set. Additionally, our T5 model achieves a character error rate (CER) of 0.77% on a set that contains real soft spelling mistakes. We achieved these results using a 4-layer T5 model trained with a 90% error injection rate, with a maximum sequence length of 300 characters.

KEYWORDS

Arabic spelling correction, Transformers, Text-to-text transfer transformer, T5, Natural-language processing.

1. INTRODUCTION

The Arabic language suffers from various types of spelling mistakes. Most of these mistakes occur due to the complex rules and the various shapes of certain letters. Soft mistakes are one of the most common spelling mistakes that deal with confusion among different shapes of certain letters. For example, the letter *alef* (ا) has two shapes (ا) and (ى) at the end of a word, like the word (عصا) which is a noun that means a stick and the word (عصى) which is a verb that means disobey. Mixing the letter *alef* in this type of words can change the meaning or make the sentence ambiguous. *Al-hamza* (ء) can be written in different shapes (ء, ا, أ, إ, ئ, ؤ) depending on the rules, like the word (قراءة) which is often incorrectly written as (قرأة). Letter *teh* (ت) can be written as (ت) or (ة) at the end of a word, like the word (ذروة); when it's indefinite and added to a definite word such as (ذروة العمل), the sound of letter *teh* indicates that it's written like this (ذروت العمل). Therefore, people tend to mix between the two shapes of the letter *teh*. Also, the letter *teh marbuta* (ة) is incorrectly written as the letter *heh* (ه) at the end of a word, like the word (ساعة) that's written often like (ساعه). Additionally, the insertion and omission of the letter *alef* (ا) after *waw aljamaea* (واو الجماعة) is also a common soft spelling mistake.

Table 1. Targeted characters and their romanization.

Characters	**Romanization**
ء	*Al-hamza*
ا	*alef*
ت	*teh*
ـه	*heh*
و	*waw*

Al-Ameri (2015) conducted a study that shows which errors are the most frequent among a group of students in a teaching institute. The number of students that participated in the study was 100 students (40 males and 60 females). In Table 2, we show the frequent errors that occurred in the study. We noticed that errors related to *Al-hamza* are the most frequent. Additionally, the errors that are related to the shape of *alef* occurred with a high percentage among the participants [1].

Awad (2012) performed a study that shows the common spelling mistakes among 130 middle-school students. As shown in Table 3, most of these errors are related to *Al-hamza*. Additionally, the insertion and omission of alef after *waw aljamaea* is the second most common error in the study [2].

M. Al-Qaraghuli and O. A. Jaafar are with Computer Center, Univ. of Baghdad, Iraq. Emails: mohammed.adel@cc.uobaghdad.edu.iq and ola.arif@cc.uobaghdad.edu.iq

Table 2. The frequent spelling mistakes as reported in Al-Ameri study.

Index	Spelling Error Type	Percentage
1	Writing *hamza mutwsita* on an *alef*	73%
2	Writing *alef maqswra* instead of *alef mamdwda*	71%
3	Writing *alef mamdwda* instead of *alef maqswra*	70%
4	Omitting *alef* following a *waw* at the end of some verb forms	67%
5	Writing *teh* instead of *teh marbuta*	67%
6	Writing *hamza mutwsita* on *waw*	64%
7	Writing *hamza* at the end of the word on *alef*	51%
8	Writing *hamza* on the line at the end of the word	47%
9	Writing *hamza* at the end of the word on *yeh*	47%
10	Dropping *lam* before the “solar letter”	38%
11	Writing *teh marbuta* instead of *teh*	37%
12	Writing *hamza mutwsita* on *yeh*	30%
13	Writing *hamza alqate* instead of *hamza Alwasl*	28%
14	Inserting *alef* after *waw* at the end of a word	25%

Table 3. The frequent spelling mistakes in Awad study.

Index	Spelling Error Type	Percentage
1	Confusing between *dād* and *dha*	60%
2	*alef* after *waw aljamaea*	59%
3	*Al-hamza Al-mutwsita*	58%
4	*Al-hamza Al-mutatarifa*	58%
5	Confusing between *teh* and *teh marbuta*	57%
6	Letters pronounced but not written	57%
7	*heh marbuta*	56%
8	Confusing between *hamza Al-wasl* and *hamza Al-qate*	54%
9	Letters written but not pronounced	52%
10	Confusing between solar and lunar *lam*	42%
11	*Al-tanwin*	44%
12	*Al-hamza Al- awilia*	23%

In Table 4, we show some of the soft spelling mistakes that occurred in Arabic company reviews written by customers [3]. We notice that most of these errors are related to *Al-hamza*, since people tend to forget to add it or use the incorrect shape of it. Additionally, we notice another common mistake which is the incorrect use of the letter (ـه) instead of the letter (ـة).

As illustrated above, soft spelling mistakes are common among various Arabic speakers regardless of age and education level and to continue the efforts to provide modern tools to help Arabic speakers and Arabic learners produce error-free text, we present in this paper a Text-to-Text Transfer Transformer (T5) model that automatically corrects Arabic soft spelling mistakes at a character level.

Table 4. Samples of soft spelling mistakes in arabic company reviews dataset.

Error type	Sentence
1. Incorrect shape of *Al-hamza*. The correct form is (أسعارهم) 2. Incorrect shape of *Al-hamza* and incorrect shape of *alef* at the end of the word. The correct form is (أغلى) 3. Incorrect shape of *Al-hamza*. The correct form is (أنصح)	اسعارهم[1] اغلا[2] من المحلات بكثير و بحطولك توصيل مجاني حكي فاضي التطبيق لا انصح[3] به
4. Incorrect shape of *Al-hamza*. The correct form is (أكثر) 5. Incorrect shape of *Al-hamza*. The correct form is (الإجراءات) 6. Using the letter (ـه) instead of the letter (ـة). The correct form is (المعاملة) 7. Using the letter (ـه) instead of the letter (ـة). The correct form is (الدقة)	برنامج اكثر[4] من رائع للتقسيط سهولة الاجراءات[5] وحسن المعامله[6] والدقه[7] واحترام العميل
8. Incorrect shape of *Al-hamza* and incorrect insertion of (ا) after the letter (و) at the end of the word. The correct form is (أرجو) 9. Incorrect shape of *Al-hamza*. The correct form is (إلى)	ارجوا[8] تشغيل خط من الحرم اليوناني الجامعة الأمريكية بوسط البلد الى[9] المقطم شارع 9

Previous works that handle Arabic soft spelling mistakes used both BiLSTM and the original transformer and the T5 model has overtaken the original transformer as the way-to-go encoder-decoder model. T5 has a simpler architecture than the original transformer, yet achieved better results compared to the other transformer models in various tasks [4][5][6]. In this paper, T5 shows a CER reduction of 10.4% over the previous works that used the original transformer and handled the same type of errors.

The rest of the paper is structured as follows: In Section 2, we will review the most recent works that are related to text correction. In Section 3, we give an overview of the T5 model. In Section 4, we show our work methodology. In Section 5, we show our results and discuss them. In Section 6, we state the limitations of our work. Finally, in Section 7, we give the conclusion and our future work ideas.

2. Literature Review

Recent works in both Arabic and foreign languages started to use transformer neural networks in text-correction tasks, whether being spelling correction, grammatical correction or post-ASR correction. In this section, we will review the most recent works that dealt with text correction for Arabic and foreign languages.

2.1 Arabic Text Correction

Al-Oyaynaa and Kotb introduced a detection system for Arabic grammatical errors using the Arabic version of the BERT model (AraBERT) and the Multilingual BERT (M-BERT). They tackled the problem at the token level and sentence level and their best results were achieved using the Arabert model with an F_1 score of 87% at the token level and an F_1 score of 98% at the sentence level [7]. Abandah *et al*. proposed an error-injection approach called stochastic error Injection to insert artificial errors into a correct text, thus providing the model with enough data to train. They used this approach to train a BiLSTM model to correct Arabic soft spelling mistakes. They achieved a CER of 1.28% on a set with real soft errors called Test200, using a 2-layer BiLSTM model with a 40% error injection rate [8]. Similar to [8], Al-Qaraghuli *et al*. introduced a transformer-based model trained from scratch to correct Arabic soft spelling mistakes. They used the original transformer architecture and trained it on a large text from Wikipedia that was injected with artificial errors using stochastic error Injection. They achieved a CER of 0.86% on the Test200 set, using a 4-layer transformer model with a 90% error injection rate [9]. Madi and d Al-Khalifa introduced three models to detect and correct a variety of Arabic errors, such as syntax errors, semantic errors and spelling errors. They achieved an $F_{0.5}$ score of 81.55% using the BiLSTM model [10].

2.2 Foreign Text Correction

Wei *et al.* proposed a detection and correction system for Chinese spelling mistakes. They built the detection part of the system based on the ELECTRA model. As for the correction part, they implemented three models based on BERT. They evaluated their system on three datasets achieving an average of 5.8% F_1 improvement over previous works [11]. Stankevičius *et al*. proposed a multilingual model based on ByT5 and statistical Unigram to correct typographical errors and restore diacritics. They achieved a 94.6% average accuracy for 13 languages [12]. Neto *et al*. introduced a spelling-correction model to recognize a handwritten text in English, French and Latin. They implemented an encoder-decoder model with the Luong attention mechanism. They achieved a CER of 3.2% and a WER of 7.7% [13].

3. Text-to-Text Transfer Transformer

Text-to-Text Transfer Transformer (T5) is an encoder-decoder transformer model proposed by Raffel *et al*. [14]. The main idea of the T5 model is to leverage transfer learning to produce a unified framework for multiple NLP tasks. T5 is built with a similar architecture to the original transformer that was proposed by Vaswani *et al*. [15]. The major difference between T5 and the original transformer is the positional encoding method, where T5 uses a simplified relative positional encoding, while the original transformer uses sinusoidal positional encoding. The other difference is that T5 removes the additive bias in the normalization layer.

T5 was pre-trained on the Colossal Clean Crawled Corpus (C4) dataset [14] and then finetuned on multiple NLP tasks. In the finetuning stage, T5 uses a prefix to identify the task that is required to work on it.

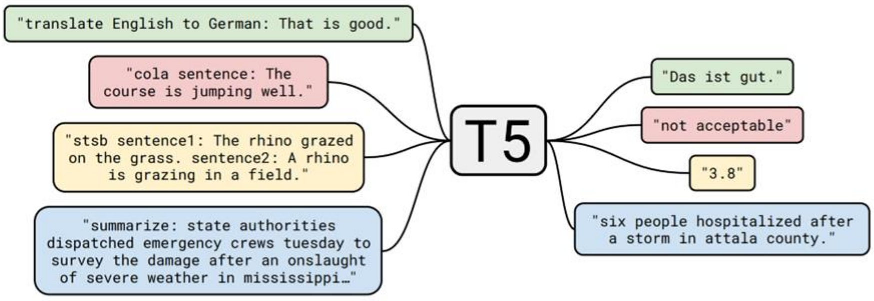

"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

Figure 1. NLP tasks and their prefixes in T5 model.

As shown in Figure 1, we can notice the prefixes such as "translate" which tells the model to translate text. T5 can translate English text into three languages; namely, German, French and Romanian. The second prefix is "summarize" which tells the model to summarize the text. The other prefixes are named based on the task's dataset such as "cola sentence" which answers whether the sentence is grammatically acceptable or not. As for "stsb sentence 1 sentence 2", it's a prefix used for sentence-similarity tasks. In this task, the model outputs a number as a string to measure whether sentence 1 is similar to sentence 2 or not.

Additionally, the authors of T5 introduced an improved version of T5 called T5 version 1.1. This version uses the GEGLU activation function instead of ReLU. Also, the embedding layer doesn't share parameters with the classifier layer.

It is worth mentioning that T5 V1.1 was only pre-trained on the C4 dataset, unlike the regular T5 which was pre-trained and finetuned on the previously mentioned tasks.

4. METHODOLOGY

In this section, we show how we trained our T5 model and the datasets we used in the training and evaluation. For training, we injected the Wiki-40B set with artificial errors and trained the T5 model on the set. Then, we finetuned our model by changing the number of layers, the error injection rate and the maximum sequence length. For evaluation, we used two sets; the Wiki-40B test set and the Test200 set. We used two evaluation metrics; BLEU Score and CER. We calculated the BLEU Score on the Wiki-40B test set and CER on Test200. Figure 2 summarizes our work methodology.

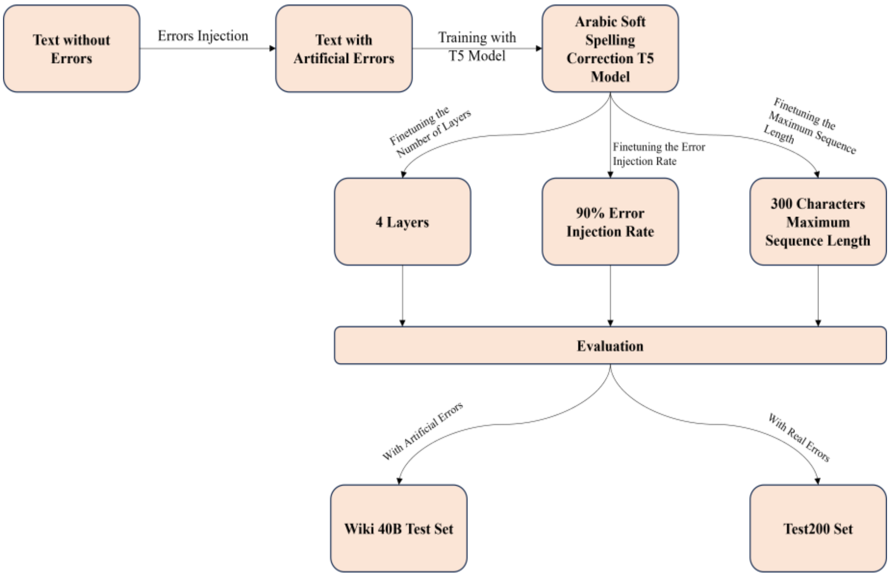

Text without Errors
Errors Injection
Text with Artificial Errors
Training with T5 Model
Arabic Soft Spelling Correction T5 Model
Finetuning the Number of Layers
Finetuning the Error Injection Rate
Finetuning the Maximum Sequence Length
4 Layers
90% Error Injection Rate
300 Characters Maximum Sequence Length
Evaluation
With Artificial Errors
With Real Errors
Wiki 40B Test Set
Test200 Set

Figure 2. The methodology of our work.

4.1 Model Building and Training

We built our T5 model using three configurations, as described in Table 5. Configuration 3 has the same parameters as the small T5 in [14].

Table 5. T5 configurations.

Parameter	Conf. 1	Conf. 2	Conf. 3
Model Dimension	128	512	512
Feed Forward Dimension	512	1024	2048
Number of Heads	8	8	8
Encoder Layers	2	4	6
Decoder Layers	2	4	6
Dropout Rate	0.1	0.1	0.1
Batch Size	128	128	64
Model Parameters	2m	21m	44m

All three configurations were trained with Kaggle TPUs. The specifications of the Kaggle platform are reported in Table 6. In the training stage, we used Adam optimizer with a learning rate = 1e-4, sparse categorical crossentropy as a loss function and used Accuracy as a training and validation metric.

In Table 7, we report our training and validation stats for the three configurations.

Table 6. Kaggle-platform specifications.

Aspect	Specification
CPU	Intel(R) Xeon(R) CPU @ 2.20GHz
TPU	v3-8
GPU	Nvidia P100 @ 1.32GHz, 16 GB
Memory	30 GB
Libraries	Python 3.10.12, TensorFlow 2.13.0

Table 7. The training and validation stats for the three configurations.

Parameter	Conf. 1	Conf. 2	Conf. 3
Number of Epochs	104	53	36
Training Time (Hours)	**21.66**	27.51	33.24
Best Accuracy on the Train Set	99.60%	99.73%	**99.75%**
Best Accuracy on the Validation Set	99.66%	**99.71%**	**99.71%**
Best Loss on the Train Set	1.29%	0.85%	**0.79%**
Best Loss on the Validation Set	1.14%	**0.99%**	1.02%

4.2 Datasets

In this work, we trained our T5 model using a dataset named Wiki-40B; for evaluation, however, we used two sets; the first one is the Wiki-40B test set (only 2k sequences were selected, because text generation will take a long time and requires using more resources) and the second one is Test200. The details of each set and preparation steps are described in the following sub-sections.

4.2.1 Wiki-40B

Wiki-40B is a multilingual dataset with 40 billion characters [16]. Wiki-40B Arabic version has 245,354 articles that were split into three sub-sets: train, validation and test.

In Table 8, we show the occurrences of the target characters in the Wiki-40B set. Character (ا) is the most occurring character with 14.27% while the character (ء) is the least occurring character with 0.04%.

Table 8. Targeted character occurrences in Wiki-40B set.

Character(s)	Wiki-40B Set
ء	0.04%
أ	1.94%
ئ	0.40%
إ	0.75%
ؤ	0.09%
ا	14.27%
آ	0.10%
ـة	3.41%
ى	0.81%
ـو	0.46%
ـت	1.16%
ـوا	0.07%
ـه	0.70%
اء	0.27%
Total	24.47%

We prepared the Wiki-40B set for our model as follows:

1. We removed punctuation marks, numbers and English letters, thus only Arabic letters remained in the set.
2. We wrapped all sequences that are longer than 300 characters, which is the selected maximum sequence length, to optimize model training.
3. Lastly, we injected artificial errors into the dataset using the stochastic error-injection approach that was proposed in [8].

Table 9. Wiki-40B transformation after being processed.

Original Text	Transformed Text
_START_ARTICLE_ احتلال الفضاء الإلكتروني _START_PARAGRAPH_ يُقصد بمصطلح احتلال الفضاء الإلكتروني (الذي يُعرف كذلك باسم احتلال النطاق الإلكتروني)، وفقًا للقانون الفيدرالي الأمريكي الذي يحمل اسم قانون حماية المستهلك ومكافحة احتلال الفضاء الإلكتروني، تسجيل اسم نطاق أو الاتجار فيه أو استخدامه بنية غير حسنة، للتربح من شهرة علامة تجارية تخص شخصًا آخر، حيث يعرض محتل الفضاء الإلكتروني بيع النطاق للشخص أو الشركة التي تمتلك العلامة التجارية الواردة ضمن الاسم بسعر مبالغ فيه_NEWLINE_.وهذا المصطلح مشتق من لفظ "الاحتلال"، الذي يعني وضع اليد على أرضٍ أو مبنى شاغر أو مهجور لا يمتلكه المحتل ولم يقم باستئجاره ولم يُمنح الإذن باستخدامه. غير أن احتلال الفضاء الإلكتروني يختلف عن ذلك إلى حدٍ ما؛ إذ يتم أحيانًا دفع مقابل أسماء النطاق التي يتم "احتلالها" عن طريق عملية التسجيل التي يقوم بها محتلو الفضاء الإلكتروني، وعادة ما يطلب محتلّي الفضاء الإلكتروني أسعارًا أعلى من تلك التي اشتروا بها. كما يضع بعض محتلو الفضاء الإلكتروني تعليقات تنتقص من شأن الشخص أو الشركة التي يعبر عنها النطاق، سعيًا لتشجيع الضحية على شراء النطاق من المحتلين، في حين يقوم آخرون بإرسال روابط مدفوعة عبر جوجل وياهو وأسك دوت كوم وغيرها من شبكات الإعلان مدفوعة الأجر، إلى الموقع الفعلي الذي كان المستخدم يريده، ومن ثم يقننون احتلالهم.	احتلال الفضاء الإلكتروني يقصد بمصطلح احتلال الفضاء الإلكتروني الذي يعرف كذلك باسم احتلال النطاق الإلكتروني وفقا للقانون الفيدرالي الأمريكي الذي يحمل اسم قانون حماية المستهلك ومكافحة احتلال الفضاء الإلكتروني تسجيل اسم نطاق أو الاتجار فيه أو استخدامه بنية غير حسنة للتربح من شهرة علامة تجارية تخص شخصا آخر، حيث يعرض محتل الفضاء الإلكتروني بيع النطاق للشخص أو الشركة التي تمتلك العلامة التجارية الواردة ضمن الاسم بسعر مبالغ فيه وهذا المصطلح مشتق من لفظ الاحتلال الذي يعني وضع اليد على أرض أو مبنى شاغر أو مهجور لا يمتلكه المحتل ولم يقم باستئجاره ولم يمنح الإذن باستخدامه. غير أن احتلال الفضاء الإلكتروني يختلف عن ذلك إلى حد ما؛ إذ يتم أحيانا دفع مقابل أسماء النطاق التي يتم احتلالها عن طريق عملية التسجيل التي يقوم بها محتلو الفضاء الإلكتروني وعادة ما يطلب محتلو الفضاء الإلكتروني أسعارا أعلى من تلك التي اشتروا بها. كما يضع بعض محتلّي الفضاء الإلكتروني تعليقات تنتقص من شأن الشخص أو الشركة التي يعبر عنها النطاق سعياً لتشجيع الضحية على شراء النطاق من المحتلين، في حين يقوم آخرون بإرسال روابط مدفوعة عبر جوجل وياهو وأسك دوت كوم وغيرها من شبكات الإعلان مدفوعة الأجر إلى الموقع الفعلي الذي كان المستخدم يريده ومن ثم يقننون احتلالهم.

In Table 9, we show the transformation of the Wiki-40B dataset after being processed. We transformed the set from long articles to short and understandable sentences. In Table 10, we show

Wiki-40B and Test200 characteristics. The dataset contains 4m sequences now which is considered a suitable size to train a transformer model.

4.2.2 Test200

We used Test200 as our evaluation set, because it contains real soft mistakes that were collected and corrected manually. The set has two hundred sequences with a total number of errors of 1,306 and with an average of 6.5 errors per sequence [8].

Table 10. Wiki-40B and Test200 characteristics.

Criterion	Wiki-40B Set	Test200 Set
Sequence count	4m	200
Word count	73.5m	2,443
Arabic letter count	354m	24,002
Words per sequence	18.3	12.2
Letters per word	4.8	9.8

4.3 Evaluation Metrics

We used two evaluation metrics; the first one is character error rate (CER) to show whether the model can correct soft spelling errors, and the lower the better the CER value.

The second metric is bilingual evaluation understudy (BLEU) to measure whether the model can produce a structured text similar to the target text [17]. The higher the better the BLEU value.

5. Results and Discussion

In this section, we show the four experiments that we conducted and the results that we obtained. We finetuned our model by changing the number of layers, using three error-injection rates and using two different maximum sequence lengths.

5.1 Model Size

In Figure 3, we show the BLEU score of the three configurations on the Wiki-40B test set. All the results are high, which indicates that all three configurations can output understandable Arabic text. We notice that configurations 2 and 3 have higher results than configuration 1, which is because the more the number of layers increased the better the model can capture the characteristics of the language and therefore, it can produce structured and understandable text. Additionally, having a high BLEU score is related to the number of targeted letters in the set. We only targeted certain letters; so, the rest of the letters should remain the same without any change. After confirming that all configurations can understand the Arabic language, now we look at CER on the Test200 set to examine which configuration is performing better.

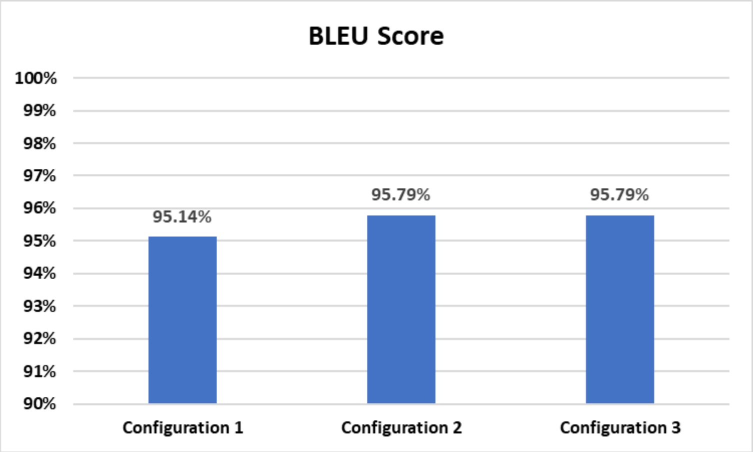


Figure 3. BLEU scores on Wiki-40B test set for the three selected configurations.

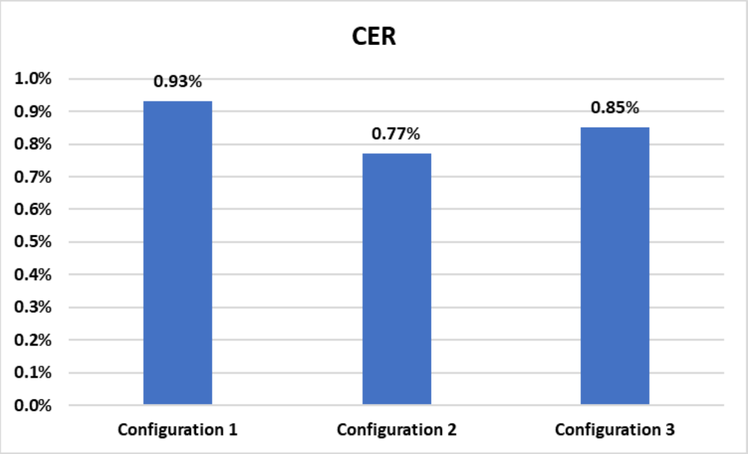


Figure 4. CER results on Test200 set for the three selected configurations.

We can observe in Figure 4 that configuration 2 achieved the lowest CER of 0.77% compared to 0.85% of Configuration 3, while configuration 1 obtained a CER of 0.93%, which is higher compared to both configurations 2 and 3. In contrast to BLEU, CER measures the model's ability to correct the errors in the set; so, the optimal value does not only depend on the number of layers. The size of the training set also plays a major factor in this process. The size of Wiki-40B is suitable for a 4-layer transformer model, as this work shows and the work in [9]. Additionally, if the size of the set is small, transformer models may not be able to outperform BiLSTM models as shown in [9]. Based on these results, we selected configuration 2 to conduct the rest of our experiments.

5.2 Error-injection Rate

In this experiment, we used three error-injection rates; 40%, 60% and 90%. In Figure 5, we see that the 90% rate is indeed the best rate for our T5 model achieving a CER of 0.77%. This result aligns with the one reported in [9] that suggests that the 90% rate is the one suitable for a transformer model, because the high injection rate gives the model more errors to correct and the attention mechanism benefits from this, in contrast to BiLSTM models that required lower injection rates, such as 40% as reported in [8].

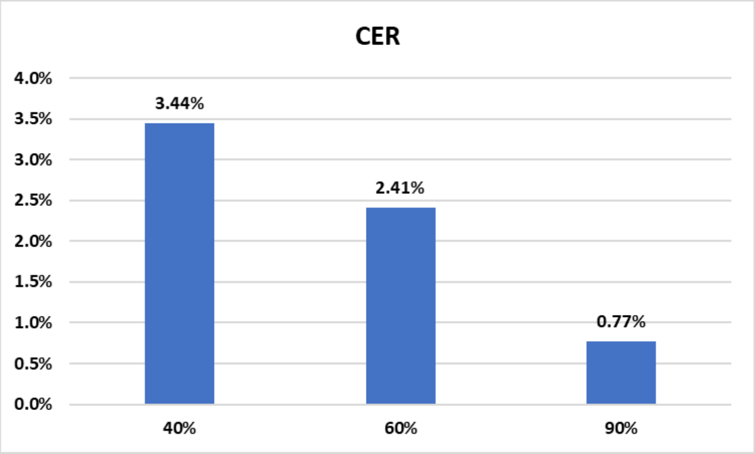


Figure 5. CER results on Test200 set using three different error-injection rates.

5.3 Maximum Sequence Length

The third experiment is related to the maximum sequence length. We tested whether increasing the maximum sequence length can improve the performance of the model. We increased the length to 500 characters; the model obtained a CER of 0.78%. The result is not improved compared to the 0.77% that was obtained using 300 characters as maximum sequence length. We also observed that training time has increased significantly when we increased the maximum sequence length, as shown in Figure

6. The model that was trained with a max. length of 300 characters took 27.51 hours, while the one trained with a max. length of 500 characters took 59.36 hours.

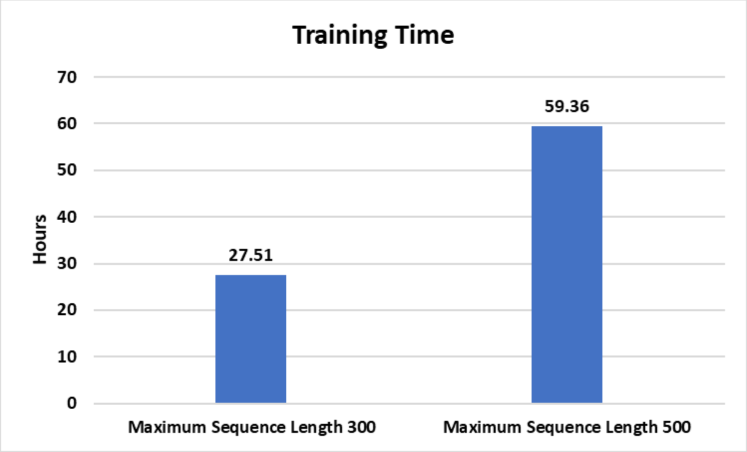


Figure 6. T5 training time in hours using 300 and 500 maximum sequence length.

5.4 T5 Version 1.1

In the final experiment, we used T5 V1.1 to check whether it could lower the results that we obtained so far. As shown in Figure 7, the results did not improve compared to the results of the regular T5.

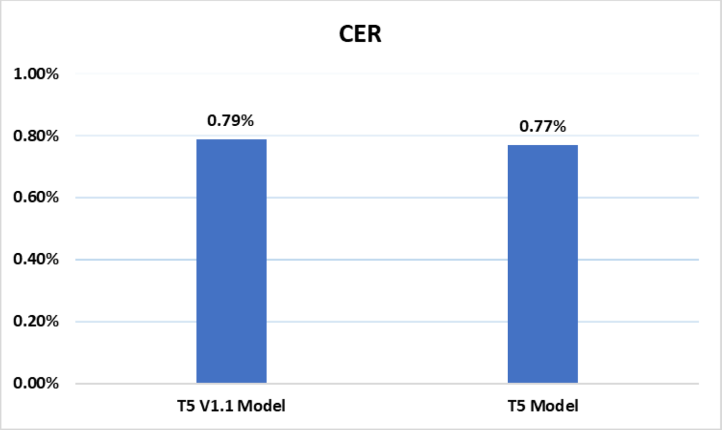


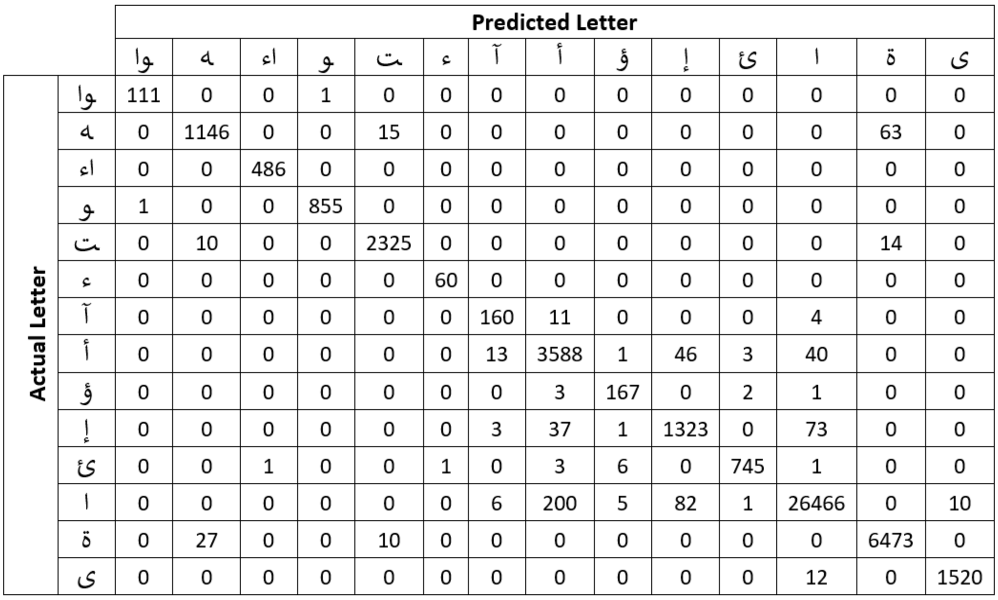
Figure 7. T5 and T5 V1.1 CER results on Test200 set.

In summary, we conducted four experiments and obtained the lowest CER of 0.77% using four layers. T5 model was trained with a 90% error-injection rate and a maximum sequence length of 300 characters.

5.5 Confusion Matrix

In Figure 8, we show the confusion matrix of our best model on the Wiki-40B test set. The set was injected with 32134 artificial errors and our model was able to correct 97.8% of these errors. We can observe that the letter (ا) is the one with the most confusion. It was falsely predicted as the letter (أ) 200 times and as the letter (إ) 82 times.

In Table 11, we show the ability of our model to correct real soft errors that were previously explained in Table 4. We can notice that the T5 model can correct the four types of errors that we targeted, which can be shown in the following words: The word (اغلا) contains two types of errors; the first one is *Al-hamza* shape at the beginning of the word and the second one is the *alef* shape at the end of the word. The model was able to correct these errors and restore the right shape of the word (أغلى). The third type of error that we targeted is the shape of *waw* at the end of a word; the word (ارجوا) contains this type of error. The model was able to correct the shape of *waw* at the end of the word (أرجو) along with *Al-hamza* shape at the beginning of the word (أرجو). The last type of error that we targeted is the shape of *teh* and *heh* at the end of a word. This error appears in two words (المعامله) and (الدقه) and both words were successfully corrected; (المعاملة) and (الدقة).



		Predicted Letter													
		او	ـه	اء	و	ـت	ء	آ	أ	ؤ	إ	ئ	ا	ة	ى
Actual Letter	او	111	0	0	1	0	0	0	0	0	0	0	0	0	0
	ـه	0	1146	0	0	15	0	0	0	0	0	0	0	63	0
	اء	0	0	486	0	0	0	0	0	0	0	0	0	0	0
	و	1	0	0	855	0	0	0	0	0	0	0	0	0	0
	ـت	0	10	0	0	2325	0	0	0	0	0	0	0	14	0
	ء	0	0	0	0	0	60	0	0	0	0	0	0	0	0
	آ	0	0	0	0	0	0	160	11	0	0	0	4	0	0
	أ	0	0	0	0	0	0	13	3588	1	46	3	40	0	0
	ؤ	0	0	0	0	0	0	0	3	167	0	2	1	0	0
	إ	0	0	0	0	0	0	3	37	1	1323	0	73	0	0
	ئ	0	0	1	0	0	1	0	3	6	0	745	1	0	0
	ا	0	0	0	0	0	0	6	200	5	82	1	26466	0	10
	ة	0	27	0	0	10	0	0	0	0	0	0	0	6473	0
	ى	0	0	0	0	0	0	0	0	0	0	0	12	0	1520

Figure 8. T5 confusion matrix on Wiki-40B test set.

Table 11. Model ability to correct soft spelling mistakes.

Input Sentence	Predicted Sentence
اسعارهم اغلا من المحلات بكثير و بحطولك توصيل مجاني حكي فاضي التطبيق لا انصح به	أسعارهم أغلى من المحلات بكثير و بحطولك توصيل مجاني حكي فاضي التطبيق لا أنصح به
برنامج اكثر من رائع للتقسيط سهولة الاجراءات وحسن المعامله والدقه واحترام العميل	برنامج أكثر من رائع للتقسيط سهولة الإجراءات وحسن المعاملة والدقة واحترام العميل
ارجوا تشغيل خط من الحرم اليوناني الجامعة الأمريكية بوسط البلد الى المقطم شارع 9	أرجو تشغيل خط من الحرم اليوناني الجامعة الأمريكية بوسط البلد إلى المقطم شارع 9

5.6 Comparison

We compared our results with those reported in [8] and [9]. These works used the same training approach and the same set of errors which is the soft errors.

In Figure 9, we can observe that our T5 model obtained the lowest CER of 0.77% compared to 0.86% of the Wiki transformer model and 1.28% of the Tashkeela BiLSTM model. The CER result shows that the T5 model outperforms both the original transformer and BiLSTM.

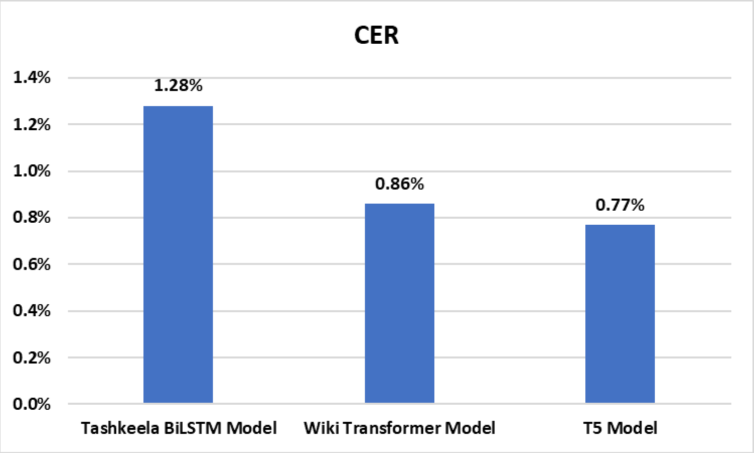


Figure 9. Our T5 CER results on Test200 set compared to the previous results of Wiki transformer model and Tashkeela BiLSTM model.

6. LIMITATIONS

In this work, we showed that the T5 model can correct Arabic soft spelling mistakes better than the original transformer and BiLSTM neural networks. Yet, this work has certain limitations, such as the lack of a large dataset that contains real soft mistakes for both training and evaluation. Additionally, using artificial errors is also limited by the size of both the dataset and the model due to the limited resources that we have. Injecting a very large set and using a large-size model are not possible in our case, requiring dedicated resources.

7. CONCLUSIONS

Nowadays, transformers dominate the natural-language processing field. Many transformer models, such as BERT, BART, T5, GPT2 and GPT3, have become the way-to-go solution for tasks, like machine translation, sentiment analysis, text generation, question answering and spelling correction. In this paper, we implemented a transformer model called T5 to correct Arabic soft spelling mistakes. We corrected four types of soft errors that are related to the shape of a letter; these types are the confusion among *Al-hamza* shapes (ء, ا, أ, إ, ئ, ؤ), the confusion among *teh*, *teh marbuta* and *heh* at the end of a word, the confusion between the two shapes of *alef* at the end of a word and the insertion and omission of *alef* after *waw aljamaea*. We achieved optimal results using a four-layer T5 model trained on Wiki-40B set that was injected with artificial errors at a 90% rate. Our model can correct 97.8% of the 32134 artificial errors that were injected into the Wiki-40B test set. We also evaluated our model using real soft errors, where our model achieved a CER of 0.77% on the Test200 set.

For future work, we would like to leverage the unified frame of T5 to create a unified model that can correct more than one type of errors, such as grammatical errors, restore and correct diacritics and other types of spelling mistakes. We are looking to use the prefix aspect of T5 to unify these types of errors and to create a training set or an evaluation set for these errors.

Additionally, we like to investigate the effects of using a bidirectional encoder model, such as BART [18] and BERT [19]. We are also looking to correct spelling mistakes at the word level using the previously mentioned models.

REFERENCES

[1] A. Al-Ameri, "Common Spelling Mistakes among Students of Teacher Education Institutes," The Islamic College University Journal, vol. 2015, no. 33, pp. 445-474, 2015.

[2] F. Awad, "Spelling Errors, Their Causes and Methods of Treatment," DIRASAT TARBAWIYA, vol. 5, no. 17, p. 217-250, 2012.

[3] Kaggle, "Arabic Company Reviews," [Online], Available: https://www.kaggle.com/datasets/fahdseddik/arabic-company-reviews.

[4] T. Adewumi, S. Sabry, N. Abid, F. Liwicki and M. Liwicki, "T5 for Hate Speech, Augmented Data and Ensemble," Sci, vol. 5, no.4, p. 37, DOI: 10.3390/sci5040037, 2023.

[5] H. Zhuang et al., "RankT5: Fine-tuning T5 for Text Ranking with Ranking Losses," Proc. of the 46th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 2308–2313, DOI: 10.1145/3539618.3592047, 2023.

[6] M. Fu et al., "VulRepair: A T5-based Automated Software Vulnerability Repair," Proc. of the 30th ACM Joint European Software Engineering Conf. and Symposium on the Foundations of Software Engineering, pp. 935-947, DOI: 10.1145/3540250.3549098, 2022.

[7] S. Aloyaynaa and Y. Kotb, "Arabic Grammatical Error Detection Using Transformers-based Pre-trained Language Models," ITM Web of Conferences, vol. 56, p. 04009, 2023.

[8] G. Abandah, A. Suyyagh and M. Z. Khedher, "Correcting Arabic Soft Spelling Mistakes Using BiLSTM-based Machine Learning," Int. J. of Advanced Computer Sci. and Appl., vol. 13, no. 5, 2022.

[9] M. Al-Qaraghuli, G. Abandah and A. Suyyagh, "Correcting Arabic Soft Spelling Mistakes Using Transformers," Proc. of the 2021 IEEE Jordan Int. Joint Conf. on Electrical Engineering and Information Technology (JEEIT), pp. 146-151, Amman, Jordan, 2021.

[10] N. Madi and H. Al-Khalifa, "Error Detection for Arabic Text Using Neural Sequence Labeling," Applied Sciences, vol. 10, no. 15, p. 5279, 2020.

[11] X. Wei, J. Huang, H. Yu and Q. Liu, "PTCSpell: Pre-trained Corrector Based on Character Shape and Pinyin for Chinese Spelling Correction," Proc. of Findings of the Association for Computational Linguistics: ACL 2023, pp. 6330–6343, Toronto, Canada, 2023.

[12] L. Stankevičius, M. Lukoševičius, J. Kapočiūtė-Dzikienė, M. Briedienė and T. Krilavičius, "Correcting Diacritics and Typos with a ByT5 Transformer Model," Applied Sciences, vol. 12, no. 5, p. 2636, 2022.

[13] A. F. de S. Neto, B. L. D. Bezerra and A. H. Toselli, "Towards the Natural Language Processing As Spelling Correction for Offline Handwritten Text Recognition Systems," Applied Sciences, vol. 10, no. 21, p. 7711, 2020.

[14] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and PJ. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-text Transformer," The Journal of Machine Learning Research, vol. 21, pp. 1-67, 2020.

[15] A. Vaswani et al., "Attention Is All You Need," Proc. of the 31st Conf. on Neural Information Processing Systems (NIPS2017), pp. 1-11, Long Beach, USA, 2017.

[16] M. Guo, Z. Dai, D. Vrandečić and R. Al-Rfou, "Wiki-40B: Multilingual Language Model Dataset," Proc. of the 12th Language Resources and Evaluation Conf., pp. 2440–2452, Marseille, France, 2020.

[17] M. Post, "A Call for Clarity in Reporting BLEU Scores," Proc. of the 3rd Conf. on Machine Translation: Research Papers, pp. 186-191, Brussels, Belgium, 2018.

[18] M. Lewis et al., "BART: Denoising Sequence-to-sequence Pre-training for Natural Language Generation, Translation and Comprehension," arXiv preprint, arXiv: 1910.13461, 2019.

[19] J. Devlin, M. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint, arXiv: 1810.04805, 2018.

ملخص البحث:

يُعدّ تصحيح أخطاء التّهجئة مهمّة تنطوي على تحدّياتٍ، وبخاصّة في اللُّغات نادرة المصادر. واللُّغة العربية هي إحدى تلك اللُّغات؛ فهي تعاني من غياب مجموعة بيانات ضخمة لتصحيح أخطاء التّهجئة. لذا تُستخدم مجموعات بياناتٍ يتمّ حقنها بأخطاء اصطناعية للتّغلّب على هذه المعضلة.

في هذه الورقة، قُمنا بتدريب محوّل النّقل من نصّ إلى نصّ (T5) باستخدام أخطاء اصطناعية لتصحيح أخطاء التّهجئة "النّاعمة" باللغة العربية. واتّضح أنّ نموذج (T5) المستخدم في هذه الدّراسة بإمكانه تصحيح ما نسبته %97.8 من الأخطاء الاصطناعية التي جرى حقنُها في مجموعة بيانات الاختبار. كذلك فقد حقّق نموذجنا معدّل خطأ أحرف (CER) مقداره %0.77 على مجموعة بياناتٍ تحتوي على أخطاء تهجئة "ناعمة" حقيقية. وقد تمّ الحصول على هذه النّتائج باستخدام نموذج (T5) رباعي الطّبقات جرى تدريبه بمعدّل حقن أخطاء مقداره %90 وطول تتابُع أقصى مقداره 300 حرف.