Arabic Misspelling Correction

خالد رشواني ، سوزانا حمزة، آية شاهين
أحمد عبود ، كنانة الغزالي، معاذ الصبح
كلية الهندسة المعلوماتية، جامعة دمشق، دمشق، سوريا

الملخص:

يهدف هذا البحث إلى تحسين تصحيح الأخطاء الإملائية والنحوية في النصوص العربية باستخدام مزيج من تقنيات التعلم العميق والخوارزميات التقليدية، تم تطبيق ثلاثة نماذج مختلفة في هذا المشروع: النموذج الأول يعتمد على AraBART، وهو نموذج seq2seq مسبق التدريب من CAMeL-Lab مصمم لتصحيح الأخطاء النحوية (Smith et al., 2021). النموذج الثاني يستخدم شبكات LSTM لمعالجة النصوص التتابعية وتحسين توقع الكلمات التالية(Hochreiter & Schmidhuber, 1997) . وأخيراً، تم استخدام خوارزمية (Levenshtein Distance) لتصحيح الأخطاء على مستوى الكلمات الفردية (Levenshtein, 1966). اعتمدت الدراسة على بيانات تصحيح الأخطاء النحوية من Hugging Face، وبيانات نصوص صحفية عربية تم جمعها من عدة مصادر، وقاعدة بيانات **Jamid** لتحليل أشكال الكلمات. أظهرت النتائج الأولية فعالية المنهجية المقترحة في تحسين جودة النصوص العربية، مما يشير إلى إمكانيات تطبيق واسعة في مجالات عدة.

الكلمات المفتاحية: (التحليل المورفولوجي، النماذج التسلسلية (Seq2Seq) ، شبكات LSTM ، تصحيح الأخطاء الإملائية)

1. المقدمة

تعد اللغة العربية واحدة من أكثر اللغات تعقيداً من حيث القواعد النحوية والصرفية، مما يجعل معالجة النصوص العربية وتصحيح الأخطاء الإملائية والنحوية تحدياً كبيراً في مجال معالجة اللغة الطبيعية (Habash, 2010). تتطلب الطبيعة المورفولوجية الغنية للغة تقنيات متقدمة يمكنها التعامل مع تعقيدات التركيب النحوي وتعدد أشكال الكلمات.

على مدى العقود الماضية، تطورت تقنيات معالجة اللغة الطبيعية، بدءاً من النماذج التقليدية المعتمدة على القواعد إلى النماذج الحديثة القائمة على التعلم العميق، مثل نماذج (Seq2Seq) وشبكات الترجمة العصبية (Bahdanau et al., 2014). في هذا البحث، نستخدم **AraBART**، وهو نموذج **(Seq2Seq)** مصمم خصيصاً للغة العربية (Moussa Kamal Eddine et al., *2022*)، وشبكات LSTM لتتبؤ النصوص التتابعية (Hochreiter & Schmidhuber, 1997) ، بالإضافة إلى خوارزمية لتصحيح الأخطاء الإملائية (Levenshtein, 1966).

تغطي الدراسة ثلاثة أنواع من البيانات: بيانات تصحيح الأخطاء النحوية من Hugging Face، وبيانات نصوص صحفية مهيأة مسبقاً، وقاعدة بيانات **Jamid** لتحليل أشكال الكلمات تهدف هذه الدراسة إلى تطوير نظام فعال وقابل للتطبيق لتحسين جودة النصوص العربية.

2. الدراسة المرجعية

2.1 Arabic Spelling Correction Using BERT (Bidirectional Encoder Representations from Transform)

هو نموذج لمعالجة اللغة الطبيعية طوره باحثو Google في عام 2018 وحقق نتائج متميزة في العديد من مهام البرمجة اللغوية العصبية. تم لاحقاً تطوير إصدارات متخصصة للغة العربية مثل Arabic-BERT، والتي استُخدمت لتحسين أداء أنظمة تصحيح الأخطاء الإملائية (Devlin et al., 2018).

تعتمد هذه الطريقة على تحويل النصوص إلى متجهات (Embeddings) لتدريب نموذج BERT. يقوم النموذج بفهم السياق واستخدام قدراته على التنبؤ بالكلمات لتصحيح الأخطاء. ومع ذلك، تشير الدراسات إلى أن دمج نموذج BERT مع نماذج أخرى يمكن أن يؤدي إلى تحسينات كبيرة في الأداء مقارنة باستخدامه منفرداً. على سبيل المثال، أظهرت التجارب التي أُجريت في جامعة Koç في إسطنبول أن الجمع بين BERT ونماذج إضافية أدى إلى تحسينات واضحة في الدقة على مستوى النصوص العربية.

Model	Arabic	Greek	Turkish	Average
SVM with TF-IDF	0.772	0.823	0.685	0.760
Multilingual BERT	0.808	0.807	0.774	0.796
Bi-LSTM	0.822	0.826	0.755	0.801
CNN-Text	0.840	0.825	0.751	0.805
BERT[4]	0.884	0.822	**0.816**	0.841
BERT-CNN (Ours)[4]	**0.897**	**0.843**	0.814	**0.851**

جدول 1 - نتائج *F1-Score* عند استخدام نماذج اخرى واستخدام *BERT* أو دمجهم .

2.2 Arabic Spell Correction with BiLSTM Model

تمثل النماذج ثنائية الاتجاه من LSTM (BiLSTM) تحسيناً لنماذج الشبكات العصبية المتكررة (RNN)، حيث إنها تعالج مشكلة فقدان السياق للكلمات البعيدة في النصوص الطويلة (Hochreiter & Schmidhuber, 1997). تعتمد هذه الطريقة على تدريب نموذج BiLSTM لفهم السياق النصي، بحيث يقوم النموذج بالتنبؤ بالكلمة التالية بناءً على السياق الحالي. يتم مقارنة تنبؤات النموذج بالكلمة الفعلية في النص:

- إذا كانت الكلمة ضمن قائمة اقتراحات النموذج، تُعتبر صحيحة.
- إذا لم تكن كذلك، يتم تصحيحها باستخدام الكلمات المقترحة ذات الصلة بالسياق.

أظهرت الدراسات أن تعزيز نموذج BiLSTM بتقنيات إضافية، مثل تحليل السياق النحوي (POS Tagging)واستخدام أسلوبn-grams ، يساهم في تحسين الأداء بشكل ملحوظ.

3. المنهجية

3.1.النماذج المستخدمة

a) AraBART : يعتمد هذا النموذج على النموذج مدرب مسبقاً AraBART (CAMeL-Lab) المُصمم خصيصًا لتصحيح الأخطاء النحوية في اللغة العربية باستخدام تقنية (Seq2Seq). تم استخدام هذا النموذج لتصحيح الأخطاء النحوية في جمل عربية تم جمعها من مجموعة بيانات Arabic Copy-GEC التي تحتوي على جمل غير مصححة ومصححة (Moussa Kamal Eddine et al., *2022*). يتم تحضير البيانات باستخدام (Tokenizer) المدمج في مكتبة Hugging Face، الذي يحول النصوص إلى تمثيلات رقمية قابلة للمعالجة بواسطة النموذج. تم تدريب النموذج باستخدام transformers مدربة مسبقاً التي تدعم

تصحيح الأخطاء النحوية (النموذج "arabart-qalb15-gec-ged-13").

b) شبكات LSTM: يتم استخدام شبكات LSTM لمعالجة النصوص التتابعية، حيث تساعد هذه الشبكات على توقع الكلمات التالية بناءً على سياق الكلمات السابقة. تم تصميم النموذج باستخدام شبكة LSTM أحادية الاتجاه ليتنبأ بالكلمة التالية في تسلسل النص، مستخدماً طبقة Embedding و طبقة LSTM لتعلم التمثيلات المورفولوجية المتقدمة للكلمات العربية (Hochreiter & Schmidhuber, 1997). تم تدريب النموذج باستخدام مجموعة بيانات نصوص صحفية عربية تم جمعها من عدة صحف (مثل الصباح وهسبريس وأخبارنا).

c) خوارزمية Levenshtein Distance : لتصحيح الأخطاء الإملائية على مستوى الكلمات الفردية، تم استخدامها حيث تتضمن هذه الخوارزمية حساب الفرق بين الكلمة الأصلية والكلمة المصححة عن طريق تحديد أقل عدد من العمليات (إدخال، حذف، أو تعديل) لتحويل كلمة إلى أخرى. تم تطبيق هذه الخوارزمية باستخدام قاعدة بيانات **jamid** التي تحتوي على كلمات وتصريفاتها المختلفة.

3.2.البيانات

i) مجموعة بيانات Arabic Copy-GEC: تتكون هذه المجموعة من جمل تحتوي على أخطاء نحوية تم تصحيحها. تم استخدامها لتدريب نموذج AraBART. حيث تم تطبيق عدة توابع تنظيف عليها ثم إضافة عمود distortion لإضافة عشوائية على النصوص بقيم مختلفة.

ii) مجموعة نصوص صحفية عربية: تتضمن هذه البيانات نصوصاً عربية من ثلاث صحف (الصباح، هسبريس، أخبارنا) تم جمعها وتحضيرها مسبقاً لاستخدامها في شبكات LSTM.

iii) قاعدة بيانات **jamid**: تتضمن هذه القاعدة أشكالاً مختلفة للكلمات العربية التي يمكن استخدامها مع خوارزمية Levenshtein Distance لتحليل الأخطاء الإملائية.

3.3.التدريب

(1) تدريب نموذج AraBART: تم تحميل النموذج المسبق التدريب "arabart-qalb15-gec-ged-13" من Hugging Face واستخدامه مع بيانات Arabic Copy-GEC لتصحيح الأخطاء النحوية.

(2) تدريب نموذج LSTM: تم استخدام شبكة LSTM لتدريب النصوص الصحفية باستخدام الدالة LSTM و طبقة Dense للتنبؤ بالكلمات التالية في النصوص التتابعية. تم تجميع التسلسلات باستخدام طبقة Embedding وتحسين التدريب باستخدام تحسين Adam و SparseCategoricalCrossentropy loss function.

(3) تطبيق Levenshtein Distance: تم تطبيق خوارزمية المسافة هذه على البيانات من قاعدة بيانات jamid لتحليل الاختلافات بين الأشكال المختلفة للكلمات وتصحيح الأخطاء الإملائية.

3.4.التقييم

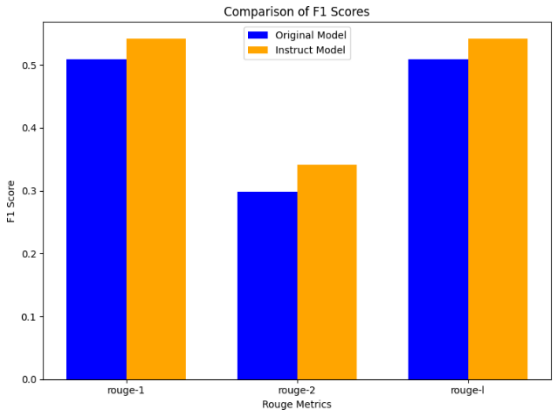
تم تقييم أداء النماذج باستخدام مقاييس مختلفة حيث تم استخدام مكتبة rouge لقياس دقة الLLM المستخدم في البداية AraBART كما تم حساب الدقة (Accuracy) لتحديد مدى فاعلية التصحيح وتحقيق النتائج المتوقعة في كل من الLSTM وخوارزمية Levenshtein. تم تقسيم البيانات إلى مجموعة تدريب و مجموعة اختبار لتقييم دقة النماذج بشكل عام.

4. التجارب والنتائج

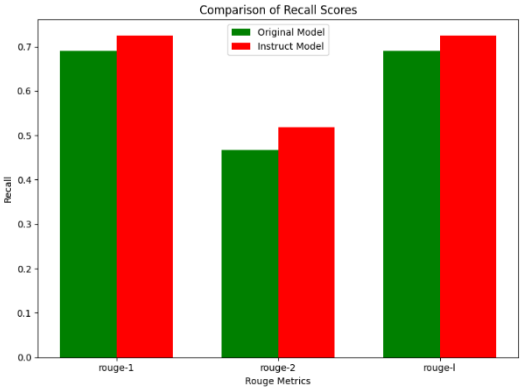
في هذه الدراسة، تم تطبيق ثلاثة نماذج لتصحيح الأخطاء الإملائية في النصوص العربية: AraBART، LSTM، وخوارزمية Levenshtein. كان المقياس الأساسي المستخدم لتقييم أداء النماذج هو الدقة، وتم مراقبة الدقة على مجموعة التحقق أثناء تدريب نموذج LSTM.

4.1 نموذج AraBART

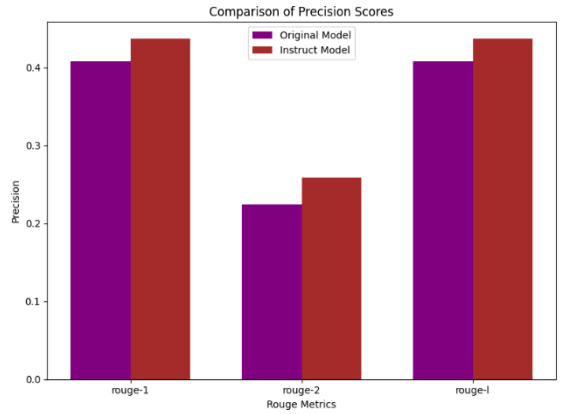
تم تقييم نموذج AraBART على مجموعة بيانات Arabic Copy-GEC، التي تحتوي على أزواج من الجمل الأصلية والمصححة. نظراً لطبيعة المهمة تم قياس الدقة من خلال مكتبة rouge ومنها حصلنا على مصفوفة قيم تحقق تظهر في الشكل 1و الشكل 2 والشكل 3 لتوضح نتائج الدقة قبل وبعد التدريب. أظهرت النتائج أن نموذج AraBART كان قادراً على تصحيح الأخطاء النحوية بدقة، مما يبرز ملاءمته لمهام تصحيح النصوص العربية.



Comparison of F1 Scores **1** *Figure*



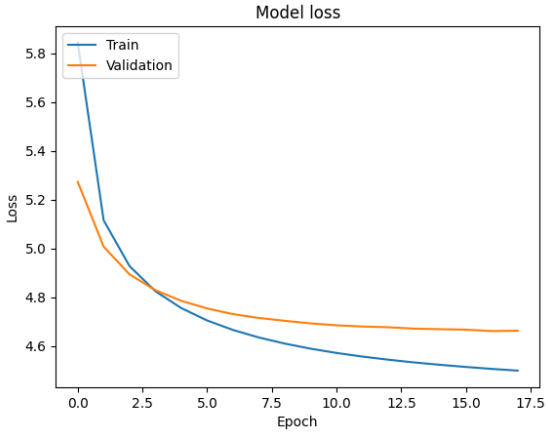
Comparison of Recall Scores **2** *Figure*



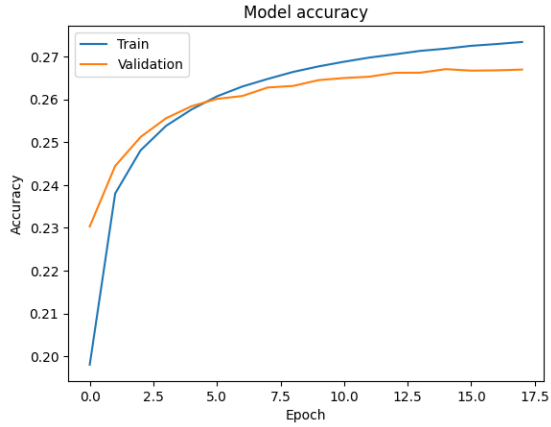
Comparison of Precision Scores **3** *Figure*

4.2 نموذج LSTM

تم تدريب نموذج LSTM على مجموعة نصوص صحفية عربية، تم معالجتها مسبقاً لإزالة الضوضاء وتوحيد التنسيق. استخدم النموذج مزيجاً من المدخلات من الجهة اليسرى وتنبؤ بالكلمات التالية في النصوص، وتم استخدام الدقة كمقياس لتقييم الأداء. حقق التدريب أداءً مرضياً مع مراقبة الدقة على مجموعة التحقق في كل فترة تدريب. تحسن أداء النموذج في التنبؤ بالكلمات التالية مع كل فترة تدريب، مما أدى إلى زيادة في الدقة على مجموعة التحقق.



LSTM Model loss **5** *Figure*



*LSTM Model accuracy***4** *Figure*

4.3 خوارزمية Levenshtein

تم تطبيق خوارزمية المسافة الليفنشتيانية لتصحيح الأخطاء الإملائية على مستوى الكلمات باستخدام مجموعة بيانات Jamid، التي تحتوي على الكلمات وأشكالها المختلفة. تم حساب المسافة بين النصوص لتقييم فعالية الخوارزمية في تصحيح الأخطاء الإملائية. رغم أن خوارزمية المسافة لا توفر دقة بالمعنى التقليدي، تم قياس فعاليتها من خلال عدد الكلمات المصححة بشكل صحيح و المسافة بين النصوص بين الكلمة المستهدفة والكلمة المصححة. أظهرت النتائج أنها كانت فعالة في تصحيح الأخطاء الإملائية في الكلمات التي تحتوي على تنوع مورفولوجي.

- **مقارنة النتائج**

أظهرت النماذج الثلاثة مزايا تكاملية:

- AraBART تميز في تصحيح الأخطاء النحوية، مما جعله الأنسب للتعامل مع تصحيح الجمل المعقدة.
- LSTM أظهر أداءً قويًا في التنبؤ بالكلمات التالية، مما ساعد في تحسين ترابط النصوص وتدفقها.
- Levenshtein قدمت تصحيحات دقيقة على مستوى الكلمات، خاصة للأخطاء الإملائية في الكلمات العربية ذات الأشكال المورفولوجية المختلفة.

من حيث الدقة العامة، تفوق نموذج AraBART على النماذج الأخرى في تصحيح الأخطاء النحوية، بينما أظهر نموذج LSTM نتائج واعدة في تحسين ترابط النصوص وتنبؤ الكلمات. كانت خوارزمية المسافة الليفنشتيانية فعالة جدًا في تصحيح الأخطاء الإملائية، ولكنها لم تتعامل مع الأخطاء النحوية.