Damascus University

Faculty of Information Technology Engineering

Grammar Error Correction (EN)

علي سيفو – قصي الشيخ علي – عامر كنهوش – غسان جربوع – بلال يونس

Supervisors:

م. زينة دلّال – م. عُلا طبّال – م. ايليسار بري – م. حاتم بركات

Abstract

يهدف هذا البحث إلى تطوير نموذج تعلم آلي قادر على تصحيح الأخطاء القواعدية في النصوص الإنجليزية باستخدام تقنيات معالجة اللغة الطبيعية (NLP). تم تنفيذ البحث عبر مرحلتين رئيسيتين:

- **المرحلة الأولى:** تضمنت بناء نموذج يعتمد على بنية Encoder-Decoder باستخدام شبكات LSTM وآلية الانتباه (Attention Mechanism) لتحسين الأداء. تم تقييم الأداء باستخدام مقياس GLEU، وأظهرت النتائج قدرة على تصحيح الأخطاء القواعدية مع بعض القيود المتعلقة بالنصوص المعقدة.
- **المرحلة الثانية:** تم استخدام نموذج (T5 Text-to-Text Transfer Transformer)، المدرب مسبقًا، الذي يُعتبر أحد أكثر النماذج تطورًا في هذا المجال، حيث أظهر النموذج أداءً متميزًا.

تضمنت الدراسة مقارنة بين نهجي Encoder-Decoder و T5، وأظهرت تفوق الأخير من حيث الأداء والكفاءة. توفر هذه النتائج أساسًا قويًا لتحسين أدوات تصحيح الأخطاء القواعدية، مع إمكانيات تطبيق واسعة النطاق

1. مقدمة

تتميز اللغة الإنجليزية بأكبر عدد من المتحدثين حول العالم. ومع ذلك، فإنها ليست اللغة الأم لغالبية هؤلاء المتحدثين. نتيجة لذلك، غالبًا ما يكون مستوى إتقانهم للغة محدودًا، مما يجعلهم أكثر عرضة لارتكاب الأخطاء القواعدية. قد تتأثر تعبيراتهم اللغوية بالتدخل من لغتهم الأم، مما يؤدي إلى أنماط أخطاء تختلف عن تلك الموجودة في كتابات المتحدثين الأصليين للغة. هذا يبرز الحاجة المتزايدة لأنظمة قادرة على تصحيح الأخطاء القواعدية تلقائيًا لمتعلمي اللغة الإنجليزية. يمكن تطبيق مثل هذه الأنظمة في سياقات متنوعة، مثل كتابة المقالات والأوراق الأكاديمية والبيانات الشخصية والأخبار ورسائل البريد الإلكتروني. ونتيجة لذلك، حظيت الأبحاث المتعلقة بتطوير أنظمة تصحيح الأخطاء القواعدية باهتمام كبير، مع تحقيق تقدم ملحوظ في السنوات الأخيرة.
يهدف تصحيح الأخطاء القواعدية (GEC) إلى تحديد وتصحيح أنواع مختلفة من الأخطاء في النص تلقائيًا. قد تكون هذه الأخطاء مخالفة لقواعد اللغة الإنجليزية أو تختلف عن الاستخدام المتوقع من قبل المتحدثين الأصليين، حيث تعمل معظم أنظمة تصحيح الأخطاء القواعدية عن طريق استقبال جملة غير صحيحة قواعدياً كمدخل وإنتاج جملة مصقولة وصحيحة قواعدياً كمخرج.

2. الدخل والخرج

2.1. مثال 1:

الدخل: "She don’t know how to do it right."

الخرج: "She doesn't know how to do it right."

هذه الجملة تحتوي على خطأ في الفعل "don’t" الذي يجب أن يكون "doesn't" لأنها تتعلق بالضمير "she".

2.2. مثال 2:

الدخل: "The books is on the table."

الخرج: "The books are on the table."

الجملة تحتوي على خطأ في استخدام الفعل "is" مع"books"، حيث يجب أن يكون الفعل في صيغة الجمع"are".

2.3. مثال 3:

الدخل: "He go to the market every day."

الخرج: "He goes to the market every day."

الخطأ في الجملة يتمثل في الفعل "go" الذي يجب أن يتوافق مع الضمير "He"في زمن الحاضر البسيط، ويجب أن يكون "goes" بدلاً من "go".

2.4. مثال 4:

الدخل: "I has finished my homework."

الخرج: "I have finished my homework."

الجملة تحتوي على خطأ في استخدام الفعل المساعد "has" مع الضمير"I" ، حيث يجب استخدام "have" بدلاً من"has" .

3. الدراسة المرجعية:

اعتمدنا في بحثنا على بعض الدراسات السابقة التي تستخدم تقنيات مشابهة لما نستعمله اليوم:

1. الورقة البحثية الأولى " Sequence to Sequence Learning with Neural Networks" **(Sutskever et al., 2014)** قدمت نموذجًا للترجمة الآلية العصبية باستخدام LSTM (Long Short-Term Memory) لتعلم ال Sequence to Sequence (Seq2Seq). النموذج المقترح يعتمد على استخدام LSTM لتشفير الجملة المدخلة إلى متجه ثابت الأبعاد، ثم استخدام LSTM آخر لفك التشفير وتوليد الجملة المترجمة، مما يعطيه القدرة على التعامل بشكل جيد مع الجمل الطويلة.

2. في الورقة البحثية الثانية **Bahdanau et al., (2014)** طُرحت آلية الانتباه (Attention Mechanism) في هذا البحث لتعزيز قدرة النماذج التسلسلية مثل LSTM في التعامل مع النصوص الأطول من خلال التركيز على الكلمات المهمة وفهم السياق بشكل أفضل، شكّل هذا العمل الأساس لدينا في بناء نموذج Encoder-Decoder LSTM with Attention

3. في الورقة البحثية Raffel et al., (2020) قُدمت الطريقة الثانية التي بنينا عليها مشروعنا اليوم فقد قدمت نموذج T5 (Text-to-Text Transfer Transformer) ، الذي يُعتبر تطويراً لنموذج Transformer استفدنا منه في فكرة تحويل كافة المهام النصية إلى صيغة Seq2Seq، ما ساعدنا على تبسيط صياغة البيانات وتحسين أداء النموذج في تصحيح الأخطاء القواعدية والحصول على نتائج جيدة جداً مقارنة بالنتائج التي حصلنا عليها بالطريقة الأولى.

4. اما الورقتين البحثيتين **(Napoles et al., 2017)** و **Lin (2004)** فقد قدمتا معيارين لتقييم جودة التصحيحات النصية وهما من اكثر المعايير انتشاراً واستعمالاً في مقارنة نتائج نماذج تحويل Seq2Seq حيث الورقة الأولى استعرضت مقياس GLEU(Generalized Language Evaluation Understanding) و الهدف منه هو تحقيق توازن بين الدقة (Precision) والتذكر (Recall)، من خلال مقارنة النصوص المصححة بالنصوص المرجعية، مع الأخذ بعين الاعتبار النصوص الأصلية التي تحتوي على الأخطاء، ما يميز GLEU هو اعتماده على مطابقة n-grams بين النص المُولَّد والنص المرجعي، مع النظر أيضًا إلى النص الأصلي. هذا يسمح له بتقييم التصحيحات البسيطة التي قد لا تكون واضحة، على سبيل المثال، إذا كانت الجملة الأصلية تحتوي على خطأ مثل: " The weather is sunny today"، وقام النموذج بتصحيحها إلى " The weather is bright today"، فإن GLEU يضع في اعتباره النص الأصلي والجملة المرجعية لتقديم تقييم دقيق للتصحيح، أما الورقة الثانية استعرضت مقياس ROUGE (Recall-Oriented Understudy for Gisting Evaluation) هذا المقياس يركز على التذكر (Recall)، أي مدى قدرة النموذج على استرجاع أكبر عدد ممكن من الكلمات أو التسلسلات (n-grams) الموجودة في النص، فهو يفضل تغطية النص المرجعي بالكامل، حتى لو كان ذلك على حساب الدقة، على سبيل المثال، إذا كانت الجملة المرجعية: " The weather in the city is wonderful"، والجملة المولدة: " The weather in the city is beautiful"، فإن ROUGE يحسب مدى تطابق الكلمات والتسلسلات بين النصين، لتقييم جودة النص المولَّد بناءً على مدى التغطية التي حققها من النص المرجعي.

4. مجموعة البيانات

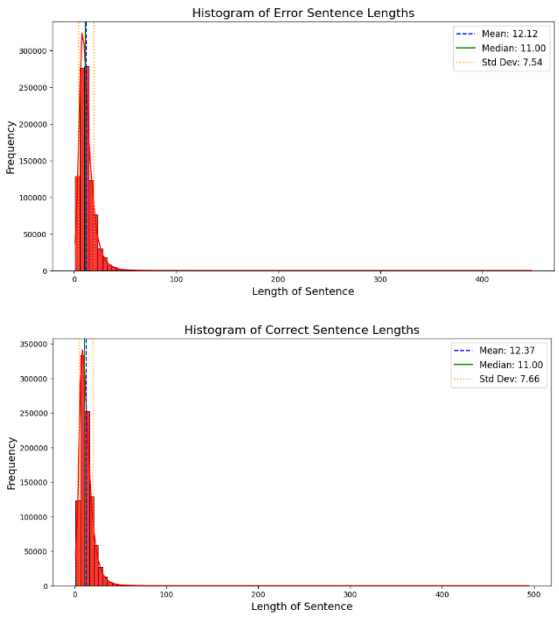
- تم استخدام بيانات GEC_LANG8 التي تتضمن نصوصًا أصلية (Input) وتصحيحاتها (Output)، توفر مجموعة البيانات هذه توازنًا بين النصوص الخاطئة والنصوص المصححة، مما يجعلها مناسبة لتدريب النماذج على اكتشاف الأنماط القواعدية الخاطئة وتصحيحها.

4.1. تحليل البيانات

- طبقنا بعض عمليات تحليل البيانات مثل طباعة ال world cloud الخاصة للكلمات الصحيحة والخاطئة لاحظنا منها الكثير من الكلمات المتشابهة ووجود الكثير من الكلمات المختصرة والاختصارات.

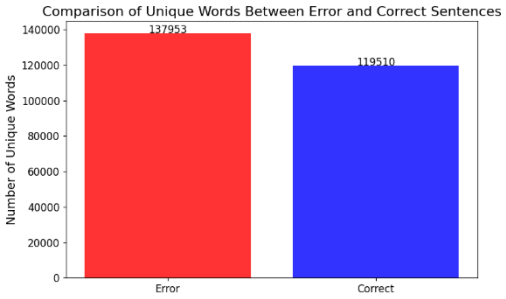
رسم توضيحي 1 - الكلمات الأكثر ظهورا في مجموعة البيانات على شكل word cloud

- رسمنا مخططات توضح أطوال الجمل الصحيحة والخاطئة لاحظنا منها ان معظم الجمل تحتوي على عدد قليل من الكلمات (10-15 كلمة) والأطوال التي تتجاوز 50 كلمة نادرة جدًا هذا يعني أن التوزيع غير متماثل ويميل نحو الجمل القصيرة.



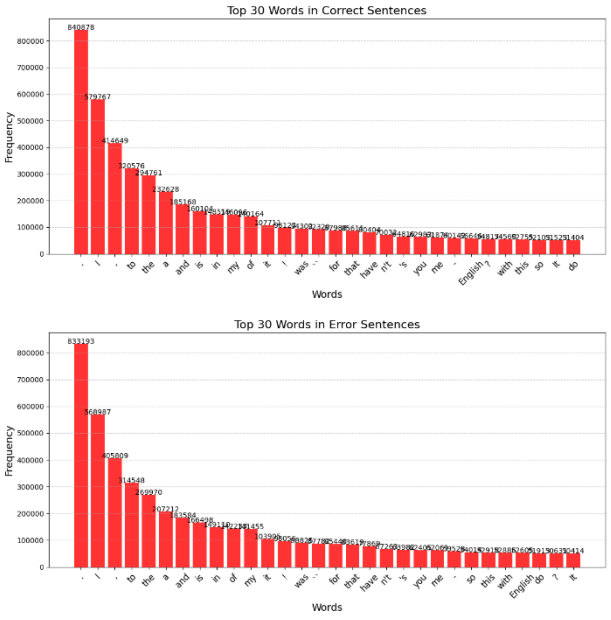
رسم توضيحي 2 –العلاقة بين تردد ظهور الكلمات في الجملة و طول الجملة الصحيحة و الخاطئة

- رسمنا مخطط يوضح عدد الكلمات الفريدة في الجمل المصححة والخاطئة، لاحظنا وجود كلمات مميزة أكثر في الجمل الخاطئة وهذا منطقي تبعاً لوجود الأخطاء.



رسم توضيحي 3 – مقارنة بين عدد الكلمات المميزة في الجمل الصحيحة و الخاطئة

- عرضنا أكثر n كلمة تكراراً في قاعدة البيانات فوجدا الكثير من علامات الترقيم والكلمات المختصرة.



رسم توضيحي 4 – الكلمات الأكثر ظهورا في كل من الجمل الصحيحة و الخاطئة

4.2. معالجة البيانات

- قمنا بتنفيذ عدة إجراءات لتحسين جودة البيانات. شملت هذه الإجراءات استبدال الكلمات المختصرة مثل "won't" بـ"will not" ، وحذف الأقواس.
- كما تمت تنقية النصوص من المحارف غير المفيدة، مع الاحتفاظ بالحروف الإنجليزية وعلامات الترقيم الأساسية مثل النقطة والفاصلة.

- تقليل التكرار في الأحرف، مثل تحويل "!!!" إلى "!"، وإزالة الأرقام تمامًا.
- تصفية العينات بناءً على طول النص، حيث تم استبعاد النصوص التي تقل عن 5 كلمات أو تزيد عن 15 كلمة لضمان توازن البيانات.

4.3. تقسيم البيانات

تم تقسيم البيانات إلى مجموعتي تدريب واختبار بنسبة 90% إلى 10% مع الحفاظ على توزيع متوازن. يضمن تقسيم البيانات وجود مجموعة اختبار تعكس الأداء الحقيقي للنموذج على بيانات جديدة.

يشير **Goodfellow et al. 2016** في كتاب "Deep Learning" إلى أهمية تقسيم البيانات للتحقق من التعميم

5. تصحيح الأغلاط القواعدية

لتصحيح الأخطاء القواعدية مقترح طريقتين لبناء النموذج، الاولى هي باستخدام ال t5 transformers والثانية باستخدام نموذج encoder-decoder.

5.1. T5 (text-to-text transfer transformer) model

- **لمحة:**

نموذج (Text-to-Text Transfer Transformer) T5 هو أحد النماذج المتقدمة في مجال معالجة اللغة الطبيعية (NLP)، وقد تم تقديمه في ورقة بحثية بعنوان "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer" بواسطة Raffel et al. (2020). يتميز هذا النموذج ببنية موحدة تعامل جميع مهام معالجة اللغة الطبيعية على أنها مشاكل تحويل Seq2Seq. بمعنى آخر، يتم تحويل كل مهمة إلى تنسيق نصي، حيث يكون الدخل والخرج عبارة عن نصوص، مما يسمح للنموذج بالتعامل مع مجموعة واسعة من المهام باستخدام إطار عمل واحد.

يتألف نموذج T5 من هيكلية مشابهة للنماذج الأخرى التي تعمل وفق نظام Encoder-Decoder. يستخدم النموذج مجموعة من طبقات المحولات (Transformer Layers) لمعالجة التسلسل النصي. يقوم الـ Encoder بمعالجة الدخل، الذي يكون عبارة عن سلسلة كلمات، باستخدام آليات Self-Attention للتعرف على العلاقات بين الكلمات والمعلومات السياقية. تسمح هذه الآلية للنموذج بالتركيز على أجزاء مختلفة من التسلسل النصي مع مراعاة الارتباطات بين الكلمات.

من ناحية أخرى، يقوم الـ Decoder بتوليد تسلسل الكلمات الناتجة خطوة بخطوة، مستخدمًا آلية Attention للتركيز على أجزاء مختلفة من الدخل أثناء التنبؤ بالكلمة التالية. تستخدم طبقات الـ Encoder في نموذج T5 آلياتMulti-Head Self-Attention ، مما يمكن النموذج من التقاط Dependencies مختلفة في الدخل والخرج. تمنح آلية Attention النموذج القدرة على التنبؤ بالكلمات حتى في الجمل الطويلة، مع التقاط المعلومات السياقية بفعالية.

يستخدم نموذج T5 أيضًا Positional Encoding لإعادة تمثيل قيم الكلمة وموقعها في الجملة، مما يساعد النموذج على فهم ترتيب وموقع الكلمات في التسلسل، وهو أمر ضروري لفهم الطبيعة التنبؤية للغة.

- **مقارنة مع GPT و BERT:**

المرجع	العيوب	المميزات	النموذج
Devlin et al., 2018	غير مناسب للتصحيح النحوي.	يعمل على تحسين فهم النصوص (Encoding) فقط دون توليد النصوص . مثالي في مهام التصنيف أو استخراج المعلومات.	Bert
Brown et al., 2020	قدرة أقل على تعلم المهام الموحدة تميزه في المهام النصية ليس بنفس مستوى T5	يركز على التوليد النصي. قوي جدًا ويمكنه أداء العديد من المهام بنجاح، بما في ذلك الترجمة.	GPT
Raffel et al., 2020	يتطلب موارد حسابية كبيرة للتدريب قد يكون معقدًا في التطبيق مقارنة بنماذج أخرى	يعتمد على تحويل النص إلى نص، مما يجعله مناسبًا لمهام متنوعة. يستخدم آلية multi-head self-attention للتعامل مع الجمل الطويلة والمعلومات السياقية. يدعم التصحيح النحوي التوليدي بشكل فعال.	T5

5.2. Encoder-Decoder LSTM with attention

1- لمحة:

يعتمد نموذج **Encoder-Decoder LSTM** مع **Attention**على استخدام شبكات الذاكرة طويلة وقصيرة الأجل (**LSTM**) لمعالجة تسلسل النصوص. يتم استخدام وحدة **Encoder**لفهم النصوص المدخلة وتحويلها إلى تمثيل مخفي(**Hidden State**) ، بينما تقوم وحدة **Decoder** بإنتاج النصوص المصححة بناءً على هذا التمثيل. يتم تعزيز النموذج باستخدام **آلية الانتباه (Attention Mechanism)** لتحسين فعالية النموذج في التعامل مع النصوص الطويلة والتعقيد اللغوي. هذا النهج تم تقديمه بشكل أساسي في عمل **(Sutskever et al., 2014)**، و لكن بدون استخدام Attention، فيما تم التطرّق لهذه الآلية في الورقة **.Bahdanau et al., (2014)**

2- البنية

- **Encoder:**

هو المسؤول عن قراءة تسلسل الكلمات في النص المدخل وتحويله إلى تمثيل مخفي (Hidden State) يلخص السياق العام للنص، و يعتمد على طبقات LSTM التي تمتاز بقدرتها على الاحتفاظ بالمعلومات الطويلة الأمد، مما يجعلها مناسبة لمعالجة النصوص الطويلة، كما يحوي الحالة المخفية النهائية (Final Hidden State) تمثل ملخصًا شاملًا للنص المدخل وتُمرر إلى Decoder.

- **Attention Mechanism:**

تعمل آلية Attention كوسيط يربط بين Encoder وDecoder بدلاً من الاعتماد فقط على الحالة النهائية للـEncoder، تقوم آلية Attention بحساب أوزان ديناميكية لكل كلمة في النص المدخل، مما يسمح للنموذج بالتركيز على أجزاء النص الأكثر صلة أثناء التصحيح.

- **Decoder:**

يستخدم الحالة المخفية الناتجة عن Encoder وأوزان Attention لتوليد النصوص المصححة خطوة بخطوة، حيث كل كلمة يتم توليدها تعتمد على الحالة السابقة والمعلومات التي تم التركيز عليها من النص المدخل.

3- آلية العمل:

- **مرحلة ال Encoding**

يتم إدخال النص المدخل كسلسلة كلمات إلى طبقات LSTM، حيث تأخذ كل خطوة زمنية كلمة واحدة وتقوم بتحديث الحالة المخفية بناءً على الكلمة الحالية والحالة السابقة. الناتج النهائي من الـ Encoderيتضمن التمثيلات المخفية لكل خطوة زمنية، التي تعبر عن سياق كل كلمة، والحالة النهائية التي تلخص جميع الكلمات.

- **آلية Attention**

لا تعتمد فقط على الحالة النهائية للـEncoder، بل تأخذ في الحسبان جميع التمثيلات المخفية. يتم حساب "درجة التركيز" لكل كلمة في النص المدخل بناءً على علاقتها بالكلمة الحالية التي يحاول الـ Decoder توليدها. يتم ذلك باستخدام معادلة تشابه، مثل Dot Product أو Scaled Dot Product، بين الحالة الحالية للـ Decoderوالتمثيلات المخفية للنص المدخل. تُحول هذه الدرجات إلى أوزان باستخدام SoftMax، مما ينتج مصفوفة أوزان Attention.

- **مرحلة Decoding**

يبدأ الـ Decoder بتوليد النص المصحّح كلمة بكلمة، حيث يتم إدخال الكلمة السابقة (أو رمز البداية في الخطوة الأولى) إلى LSTM مع الحالة المخفية. تُدمج مخرجات LSTM مع أوزان Attention لتحديد المعلومات الأكثر صلة من النص المدخل، ثم يُمرر الناتج إلى طبقة إخراج مع الدالة SoftMax للتنبؤ بالكلمة التالية. تتكرر هذه العملية حتى يتم توليد النص المصحح بالكامل أو الوصول إلى رمز النهاية.

6. التحديات

واجهنا خلال العمل على نموذج **Encoder-Decoder LSTM** مع **Attention** عدة تحديات أثرت على الأداء النهائي. أول مشكلة لاحظناها كانت فقدان النموذج للسياق في النصوص الطويلة أو المعقدة، حيث إن قدرة LSTM على التعامل مع المعلومات بعيدة المدى ليست بالقوة المطلوبة، حتى مع استخدام آلية Attention. بالإضافة إلى ذلك، وجدنا أن عملية التدريب كانت بطيئة نسبيًا بسبب الطبيعة التسلسلية لـ LSTM، مما جعل تدريب النموذج على مجموعات بيانات كبيرة يتطلب وقتًا وجهدًا أكبر بغض النظر عن عتاديات أجهزتنا المتواضعة، فقمنا بحصر التدريب على الجمل التي يتراوح طولها بين ال 5 إلى 15 كلمة. وأخيرًا، لوحظ أن النموذج يواجه صعوبة في التعميم على نصوص جديدة تختلف عن تلك المستخدمة أثناء التدريب، مما يعني أن أداءه يقل عند مواجهة أنماط أخطاء لم يعتد عليها.

7. التقييم

لتحديد كفاءة النموذج المدرب يجب تقييم خرج هذا النموذج، معيار التقييم العام عند التعامل مع تصحيح الأغلاط القواعدية هو مدى قدرة النموذج على تصحيح الجمل الخاطئة. يمكن تقسيم طرق التقييم في GEC إلى مرجعية (reference-based) وغير مرجعية (reference-less) والاختلاف بينهما هو إذا كان المرجع موجودا عند التقييم استخدمت الطرق المرجعية في هذا العمل لأن مجموعة البيانات تتألف من جملة خاطئة وجملة صحيحة ونريد تقييم الجملة الناتجة عن النموذج استنادا إلى الجملة المرجعية الصحيحة. الطريقتان المستخدمتان للتقييم هما GLEU وROUGE.

8. النتائج:

1- T5:

الطريقة	عدد العينات	Epochs	وقت التدريب	طريقة التدريب	GLUE	ROUGE 1	ROUGE 2	ROUGE L
T5	200000	1	1h	Google Colab gpu	0.5751	0.8043	0.6333	0.7918

2- Encoder-Decoder:

الطريقة	عدد العينات	Epochs	وقت التدريب	طريقة التدريب	F-score	loss
Enc-dec	10000	11	44h	Own cpu	0.6112	0.2776

3- الاستنتاج:

بناءً على النتائج المستخلصة من التجارب، يتضح أن نموذج T5 يتفوق بشكل كبير على نموذج Encoder-Decoder في معظم الجوانب المهمة لتصحيح الأخطاء القواعدية في النصوص الإنجليزية. أظهر نموذج T5 أداءً متميزًا من حيث المقاييس الدلالية مثل GLEU وROUGE (1, 2, L)، حيث حقق مستويات عالية من الدقة في تصحيح الأخطاء مقارنة بالنموذج التقليديEncoder-Decoder .

كما أثبت نموذج T5 كفاءته العالية في استغلال الموارد الزمنية والحسابية، حيث أكمل عملية التدريب في وقت قصير جدًا(ساعة واحدة) باستخدام وحدة معالجة الرسوميات GPU مع الاستفادة من مجموعة بيانات كبيرة تضم 200,000 عينة. على العكس، تطلب نموذج Encoder-Decoder وقتًا طويلًا جدًا للتدريب (44 ساعة) باستخدام وحدة المعالجة المركزية.

على الرغم من أن نموذج Encoder-Decoder أظهر قيمة خسارة منخفضة خلال التدريب، إلا أن ضعف مقاييس F-Score يدل على أنه أقل كفاءة في معالجة وتصحيح النصوص مقارنةً بنموذج T5. يدل هذا الفارق في الأداء إلى البنية المعمارية المتقدمة لنموذج T5، والتي

تعتمد على محولات (Transformers) قوية، مما يمنحه ميزة كبيرة في فهم السياق اللغوي وتصحيحه بدقة.

بناءً على هذه النتائج، يمكن القول إن نموذج T5 يمثل خيارًا ممتازاً لتطبيقات تصحيح الأخطاء القواعدية في النصوص، لاسيما في بيئات تمتلك موارد حاسوبية قوية وقادرة على التعامل مع نماذج كبيرة. في الوقت نفسه، تُبرز هذه النتائج أهمية استخدام مجموعات بيانات أكبر وتقنيات أكثر تقدمًا لتحسين أداء النماذج التقليدية مثل Encoder-Decoder.

References

[1] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*, 2014.

[2] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[3] Brown, T., et al. (2020). Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165.*

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[5] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL Workshop on Text Summarization* (pp. 74-81).

[6] Napoles, C., et al. (2017). GLEU: Generalized language evaluation understanding. *Journal of Language Technology, 28*(2), 123-145.

[7] Raffel, C., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(1), 1-46.

[8] Wang, Y., Wang, Y., Liu, J., & Liu, Z. (2020). A comprehensive survey of grammar error correction. *International Journal of Linguistics, 15*(3), 45-67.