



جامعة دمشق

كلية الهندسة المعلوماتية



## معالجة اللغات الطبيعية

### English Grammar Error Correction

تقرير إنجاز المشروع

إشراف:

م. زينة الدلال - م. علا طبال - م. ايليسار بري.

تقدمة:

حمزة محمد المحروس

أسامه يوسف بازو

ياسين راتب عبد المهدي

عليما ماجد المسوتي

28-12-2023

1	مقدمة	2
1.1	الدافع	2
2.1	هيكل العمل	2
2	الأغلاط القواعدية و تصحيحها	2
1.2	أنواع الاغلاط القواعدية	2
2.2	الدراسة المرجعية	3
2.2.1	دراسة موسعة في بحث:	4
	Neural Grammatical Error Correction Systems with Unsupervised Pre-training on Synthetic Data	4
	confusion sets	4
	النتائج	5
3.2	القيود	6
3	مجموعة البيانات	7
3.1	Exploratory data analysis	7
3.1.1	Word Cloud	7
3.1.2	Histogram	7
3.2	معالجة البيانات	9
4	تصحيح الأغلاط القواعدية باستخدام الـ transformers	10
1.4	نموذج T5 transformers	10
2.4	Encoder Decoder and Attention Model: نموذج	10
5	النتائج و التقييم	11
1.5	مقياس التقييم	11
	Rouge score	11
	GLEU	12
2.5	النتائج	12
	نتائج الطريقة الأولى	12
	نتائج الطريقة الثانية (Encoder Decoder model with attention):	12
6	الملخص	13
7	المراجع	14

# 1 مقدمة

## 1.1 الدافع

تمتلك اللغة الانجليزية أكبر عدد من المتحدثين حول العالم. نسبة كبيرة من المتحدثين لديهم اللغة الاساسية مختلفة عن الإنجليزية و يتحدثون اللغة الانجليزية بصفتها اللغة الثانية، وبالتالي لديهم مستوى الإتقان للغة غير كافي و يميلون إلى ارتكاب الاغلاط النحوية بشكل متكرر. لهذا السبب أصبح بناء نظام لتصحيح الاغلاط القواعدية تلقائياً أمراً ضرورياً في الأونة الاخيرة، و يمكن تطبيق النظام في العديد من السيناريوهات التي تفيد متحدثي اللغة الانجليزية كاستخدامه كمساعد ألي عند كتابة المقالات و الأبحاث و البيانات و الاخبار و رسائل البريد الإلكتروني [1]. تهدف أنظمة تصحيح الأخطاء النحوية Grammatical Error Correction اختصاراً (GEC) لكشف وتصحيح الأخطاء النحوية في جملة معينة. هناك خطان للعمل لبناء GEC. الطريقة الاولى sequence to sequence methods حيث يكون الدخل جملة خاطئة و يتم توليد جملة خالية من الأغلاط بشكل تلقائي، أما الطريقة الثانية sequence labeling methods تحول المطلوب إلى إيجاد سلسلة من عمليات التعديل على النص و ثم تستخدم sequence labeling للنتيئة بهذه العمليات [2]. نتطرق في هذا المشروع إلى استخدام وتقييم اداء الطريقة الاولى بالاستعانة ب transformers و ذلك في كشف وتصحيح الأغلاط القواعدية.

## 2.1 هيكل العمل

هذا المشروع يبحث في إمكانية استخدام طرائق الذكاء الصناعي في تصحيح الأغلاط القواعدية تلقائياً. نتحدث في القسم الثاني عن أنواع الأغلاط القواعدية و كيفية تصحيحها ثم ننتقل إلى توصيف مجموعة البيانات المستخدمة و ذلك في القسم الثالث. القسم الرابع نتحدث عن استخدام ال transformers في اكتشاف وتصحيح الأغلاط وفي النهاية نقيم أداء النموذج في القسم الخامس.

## 2 الأغلاط القواعدية و تصحيحها

### 1.2 أنواع الاغلاط القواعدية

الاغلاط القواعدية هي الأغلاط التي تنتهك قواعد اللغة الانجليزية ويمكن أن تصنف إلى عدم التوافق بين الفعل و الفاعل، استخدام الزمن الخاطئ للفعل، الاستخدام الخاطئ للصفات أو الظروف، الاستخدام الخاطئ للكلمات، الاستخدام الخاطئ للفاصلة العليا، الفواصل الناقصة، خلط الكلمات المتشابهة، خلاف في الضمائر المستخدمة، المقارنة الخاطئة و الاستخدام الخاطئ لحروف الجر. الجدول (1) يوضح الاغلاط من خلال الامثلة.

نوع الغلط	الجملة الخاطئة	الجملة الصحيحة
Subject-verb disagreement	People <u>is</u> coming to my party.	People <u>are</u> coming to my party.
Wrong tense	I <u>have been</u> to Damascus last summer.	I <u>went</u> to Damascus last summer.
Misusing Adverbs-Adjectives	I want to speak English <u>good</u> .	I want to speak English <u>well</u> .
Wrong use of words	I must <u>to buy</u> a new notebook.	I must <u>buy</u> a new notebook.
Apostrophe Usage	It is my <u>brothers</u> house in Syria.	It is my <u>brother's</u> house in Syria.
Missing Coma	Ahmed came to my house and Hamza joined him.	Ahmed came to my house, and Hamza joined him.
Mixing up similar words	The book has a good <u>affect</u> on my mood.	The book has a good <u>effect</u> on my mood.
Pronoun Disagreement	Every girl must bring <u>their</u> books to school.	Every girl must bring <u>her</u> books to school
Comparison	She is <u>more taller</u> .	She is <u>taller</u> .
Prepositions	I went to school <u>at</u> Sunday.	I went to school <u>on</u> Sunday.

الجدول (1)

## 2.2 الدراسة المرجعية

كان أول ظهور لأنظمة GEC في بداية الـ 2000 [1]. كانت أغلب الأنظمة في البداية تعتمد على استخدام المحلل اللغوي و القواعد اللغوية في إنشاء قواعد تكشف وتصحح الأخطاء النحوية كما في الورقة البحثية [4]. ومع ذلك، فإن تعقيد تصميم القواعد وحل النزاعات بين القواعد يتطلب قدرًا كبيرًا من العمل. على الرغم من أن بعض أعمال GEC اليوم لا تزال تستخدم القواعد كمصدر إضافي للتصحيح، فقد تم استبدال أداء أنظمة GEC القائمة على القواعد بنهج يعتمد على البيانات [1]. ظهرت أساليب تعتمد على التصنيف في أواخر الـ 2000 حيث درب النموذج على عدد كبير من الجمل الصحيحة قواعديا ليتم من خلاله التنبؤ بالكلمة التالية الصحيحة نحويًا بالاعتماد على الكلمات السابقة. في أواخر الـ 2010 بنيت أنظمة GEC تعتمد على (statistical machine translation (SMT، فتم التعامل مع المسألة كمسألة ترجمة و تم تدريب الـ SMT model على ثنائيات من الجمل (خاطئة، صحيحة) و يقوم الـ model بتصحيح الأخطاء عن طريق "ترجمة" الجملة المغلوطة نحويًا إلى جملة صحيحة، و بعدها ظهرت أنظمة GEC القائمة على (neural machine translation (NMT والتي تطبق نماذج seq2seq. في هذا المشروع قمنا ببناء نظام GEC قائم على NMT.

## 2.2.1 دراسة موسعة في بحث:

### Neural Grammatical Error Correction Systems with Unsupervised Pre-training on Synthetic Data

يعالج هذا البحث مشكلة ندرة البيانات التي قد يتم مواجهتها عند القيام بمعالجة الأخطاء القواعدية للغة، حيث تعتبر هذه الطريقة التي تندرج تحت الـ unsupervised pre-training فعالة وذلك لاستنادها إلى الـ confusion sets والتي تمكن من زيادة كمية البيانات. تستخدم هذه البيانات المولدة لتدريب pre-train Transformer sequence-to-sequence، مما يُحسن الأداء مقارنةً بالأسس القوية المدربة على بيانات أخطاء حقيقية [8].

#### confusion sets

هي مجموعات من الكلمات التي تكون متشابهة في الكتابة أو النطق ولكن لها معاني مختلفة، مما يُمكن أن يؤدي إلى الخطأ في الكتابة أو الفهم. في سياق تصحيح الأخطاء النحوية يُستخدم هذا المفهوم لتوليد أخطاء اصطناعية بهدف تدريب النماذج. حيث تشمل كلمات مقترحة كبداية محتملة لكلمة معينة. هذا يُمكن أن يُحاكي الأخطاء الشائعة التي يمكن أن يرتكبها المستخدمون الحقيقيون، ويُساعد في تدريب نماذج أكثر فعالية لتصحيح الأخطاء. الجدول (2) يعطي مثال عن هذه المجموعات.

Word	Confusion set
has	Haas HS Hans hats gas had Ha ha As as
is	IRS ISO OS US us Si its
island	islands Iceland slant
issued	issues issue used issuers eased sued assumed assured missed
student	students strident stunt
walking	talking whaling
large	larger lag lake barge Lodge lodge
largest	latest longest

الجدول (2)

قام البحث بعدها بتوليد بيانات بالاستعانة بالـ confusion set لتصحيح الأخطاء النحوية عن طريق استبدال كلمات عشوائية في جملة خالية من الأخطاء باستخدام زوج من الجمل الاصطناعية والأصلية كمثال تدريبي جديد. في نهج بسيط، يمكن استبدال الكلمات عشوائيًا ضمن المفردات، لكن هذا قد يؤدي إلى أنماط أخطاء غير واقعية لا تشبه تلك الموجودة في البيانات الحقيقية. يمكن توليد أخطاء أكثر دقة عن طريق استبدال الكلمات فقط ضمن confusion set التي تتكون من كلمات يتم الخلط بينها عادة.

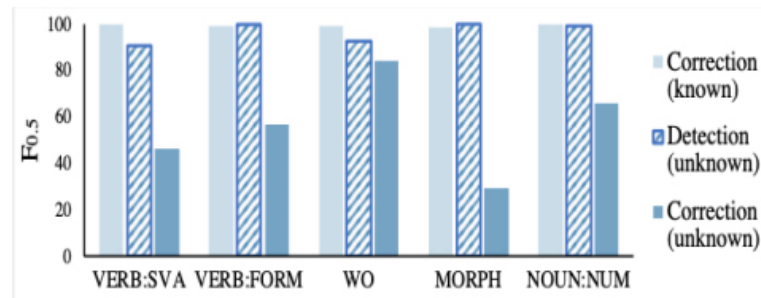
يقوم هذا البحث باستخدام نماذج تصحيح الأخطاء النحوية المبنية على Encoder-Decoder-based GEC models، التي لا تقتصر على اكتشاف الأخطاء فحسب، بل تتطلب أيضًا القدرة على تصحيحها على عكس الدراسات الأخرى التي تناولت قدرات neural language models في معرفة الصحة النحوية في مهام مثل الفاعل والفعل وغيرها في الجمل.

لتقييم قدرة نماذج تصحيح الأخطاء النحوية على التعميم، مقارنةً بين تصحيح الأخطاء المعروفة سابقًا وتصحيح الأخطاء غير المعروفة من قبل يتم استخدام بيانات مولدة وبيانات حقيقية للتقييم، ويستكشف البحث خمسة أنواع من الأخطاء المعتمدة على قواعد محددة.

Dataset		VERB:SVA	VERB:FORM	WO	MORPH	NOUN:NUM
Synthetic data	Known	99.61	99.17	99.09	98.44	97.47
	Unknown	46.05	56.93	84.00	29.35	65.55
	$\Delta$	<b>-53.56</b>	<b>-42.24</b>	<b>-15.09</b>	<b>-69.09</b>	<b>-31.92</b>
Real data	Known	87.84	86.36	74.89	87.77	83.75
	Unknown	6.28	6.28	9.25	3.83	12.49
	$\Delta$	<b>-81.56</b>	<b>-80.08</b>	<b>-65.64</b>	<b>-83.94</b>	<b>-71.26</b>

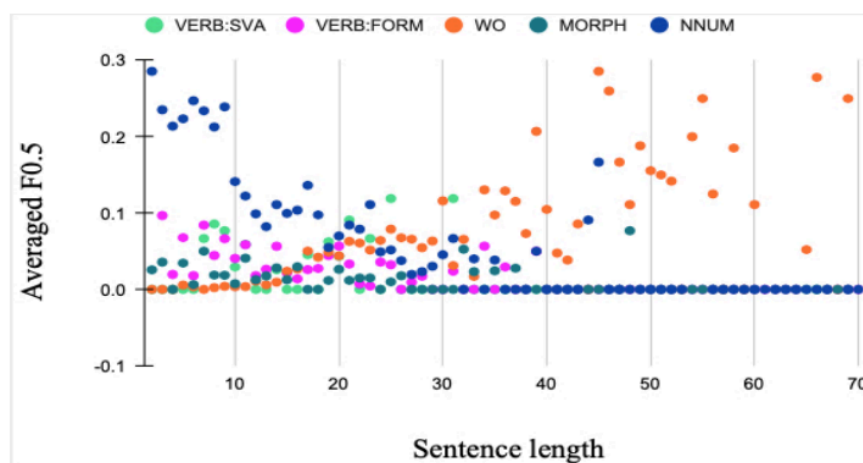
الجدول (3)

الجدول (3) يقدم أداء قدرة التعميم للأخطاء غير المعروفة من قبل، حيث يتم تقديم كل رقم كدرجة F0.5. هذه الدرجة تعكس مدى فعالية النماذج في تصحيح الأخطاء النحوية التي لم ترد في البيانات التدريبية. تم استخدام هذا التقييم لقياس القدرة على التعميم النحوي للنماذج في مواجهة أنواع مختلفة من الأخطاء. تظهر نتائج التقييم باستخدام البيانات التي تم توليدها انخفاضاً كبيراً في أداء تصحيح النموذج في unknown data مقارنة بال known data، أما باستخدام البيانات الحقيقية، فهناك انخفاض ملحوظ في الأداء على جميع الأخطاء، في unknown data، مما يشير إلى صعوبة أكبر في التعميم في سياقات أكثر واقعية.



المخطط (1)

بالنظر إلى القدرة على اكتشاف الأخطاء مقابل تصحيحها. تم قياس أداء الاكتشاف وفق ما إذا كان النموذج يجري أي تعديل في موقع الخطأ. أظهرت النتائج أن النموذج يكتشف جميع أنواع الأخطاء بنجاح، مما يشير إلى قدرته على التعميم النحوي لكشف الأخطاء، لكن ليس بما يكفي للتنبؤ بالكلمة الصحيحة.



المخطط (2)

تم تحليل تأثير عاملين: F0.5 score وطول الجملة. تم مقارنة الأداء عندما يكون الخطأ المستهدف هو الخطأ الوحيد في الجملة. كما تم التجربة في تأثير إضافة عدد قليل من أنماط تصحيح الأخطاء إلى بيانات التدريب، وأظهرت النتائج أن إضافة حتى عينة أو اثنتين يمكن أن تحسن أداء النموذج بشكل كبير.

#seen patterns	0	1	2
Precision	43.31	47.16	57.65
Recall	47.92	52.52	63.70
F <sub>0.5</sub>	44.16	48.14	58.77

#### الجدول (4)

تُظهر النتائج أن هناك تحديات في تعميم تصحيح الأخطاء على البيانات، خاصةً في البيانات الحقيقية. كما يكتشف النموذج الأخطاء بنجاح، لكنه قد لا يكون دائمًا دقيقًا في التنبؤ بالكلمة الصحيحة.

تمت مقارنة النتائج باستخدام الـ F0.5 score ومعرفة تأثير طول الجملة على أداء النموذج أي هناك علاقة بين طول الجملة واداء التصحيح.

أظهرت النتائج أيضاً أن إضافة حتى عدد قليل من أنماط تصحيح الأخطاء إلى بيانات التدريب يمكن أن تحسن الأداء بشكل كبير.

### 3.2 القيود

تتمتع مسألة GEC عند استخدام الشبكات العصبونية بتحديات و هما data sparsity و multi-pass decoding [3]. تشير مشكلة data sparsity إلى عدم وجود عدد كافٍ من أزواج الجمل الخاطئة والصحيحة لتدريب نماذج GEC القوية، حيث يكون إنتاج بيانات عالية الجودة و قريبة للواقع مرفقة بتصحيحاتها محدوداً أو مكلفاً. يمكن أن تؤدي هذه الندرة إلى نماذج أقل قابلية للتعميم أو ذات أداء ضعيف في أنواع الأخطاء الأقل شيوعاً. multi-pass decoding يعني أنه قد يتم تصحيح الجملة في عدة "تمريرات" أو جولات، مما يؤدي تدريجياً إلى إصلاح الأخطاء التي لم يتم تصحيحها في الجولات السابقة أو التي تم تقديمها من خلال التصحيحات السابقة. تواجه مسألة GEC أيضاً بعض التحديات بسبب التغيرات غير المحدودة باللغة، فعلى الرغم من وجود تصنيف للأخطاء القواعدية فإن بناء نموذج قادر على تصحيح جميع الاغلاط الممكنة هو أمر صعب خصوصاً عندما يكون شكل الغلط غير موجود في عينة التدريب [1].

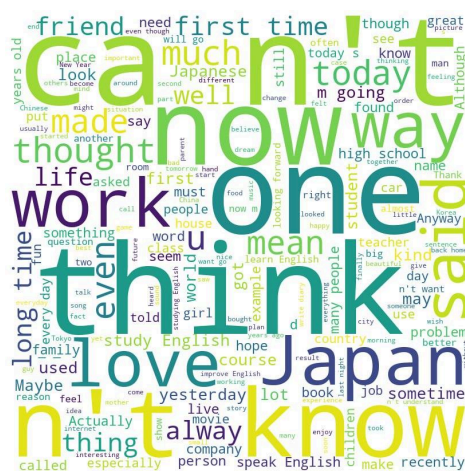
### 3 مجموعة البيانات

اعتمدنا في المشروع على مجموعة البيانات (lang-8 dataset)، البيانات عبارة عن مجموعة من النصوص التي تم تجميعها من منصة lang-8 وهي عبارة عن منصة لتعلم اللغات. تعد هذه البيانات شائعة الاستخدام في مجال التصحيح القواعدي للغة نظرا لاحتوائها على عينات حقيقية من المتعلمين بالإضافة إلى اللغات المختلفة التي يمكن الحصول عليها. تتألف البيانات من عمودين، يمثل العمود الأول النص الخاطئ، بينما يمثل العمود الثاني النص وفق التصحيح القواعدي الصحيح للجملة.

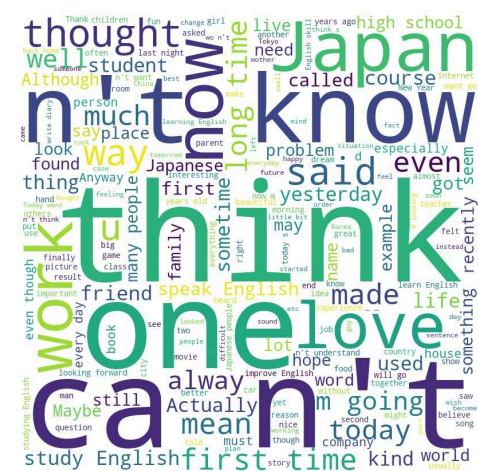
## Exploratory data analysis 3.1

طبقنا بعض العمليات التحليلية على مجموعة البيانات قبل البدء في تنظيفها فكانت النتائج كالتالي:

### Word Cloud 3.1.1



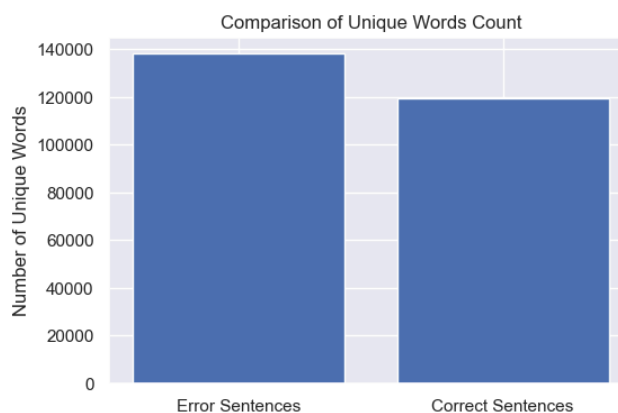
المخطط (4)  
النصوص الخاطئة



المخطط (3)  
النصوص الصحبة

نلاحظ من الصورتين وجود الكثير من الكلمات المتشابهة بين النصين، توجد الكلمة المختصرة n't بشكل واضح بين النصين فيمكن الاستفادة من هذه المعلومات لحذف الكلمات المختصرة والاستعاضة عنها بالكلمة كاملة مثل not.

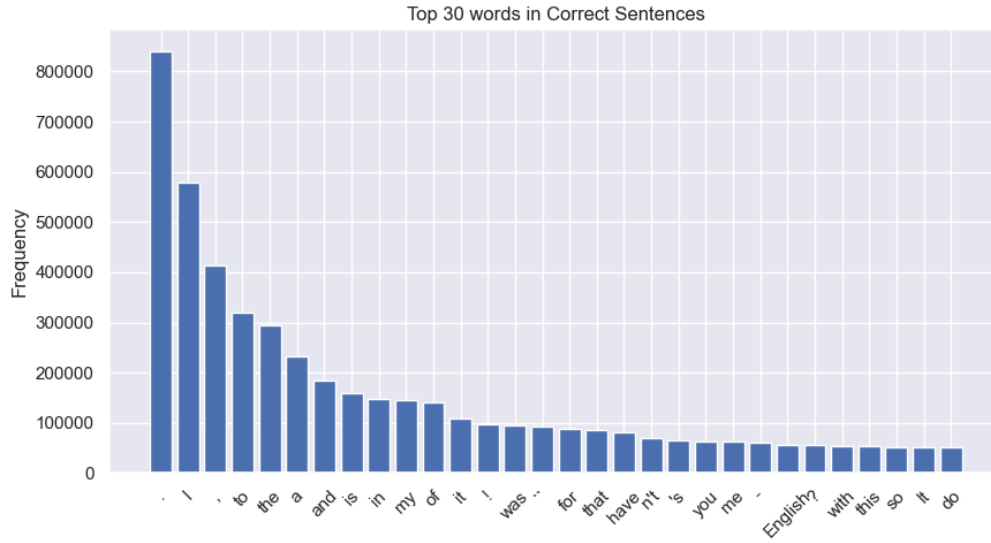
### Histogram 3.1.2



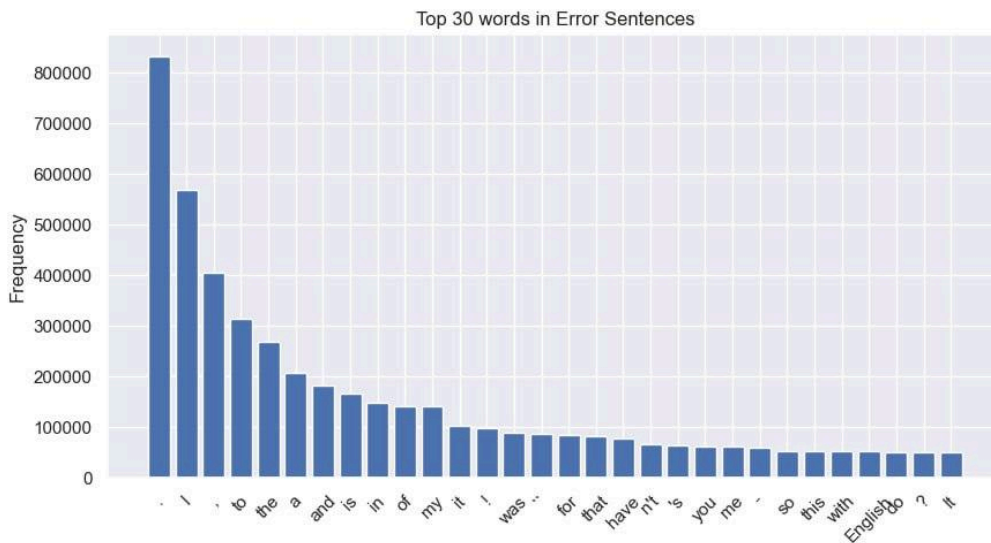
المخطط (5)

يمثل المخطط السابق عدد الكلمات الفريدة في كل من الجمل الخاطئة والجمل الصحيحة قواعدياً، تتميز الجملة الخاطئة بعدد كلمات مميزة أكثر من الجمل الصحيحة وهذا الأمر منطقي تبعاً لوجود الأخطاء في الجمل مما يؤدي لزيادة عدد الكلمات الفريدة.



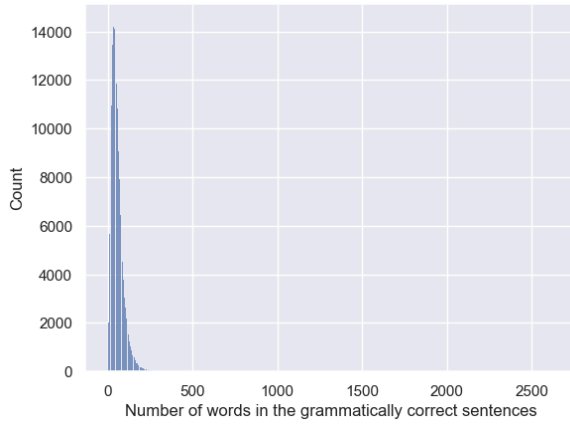


المخطط (6)

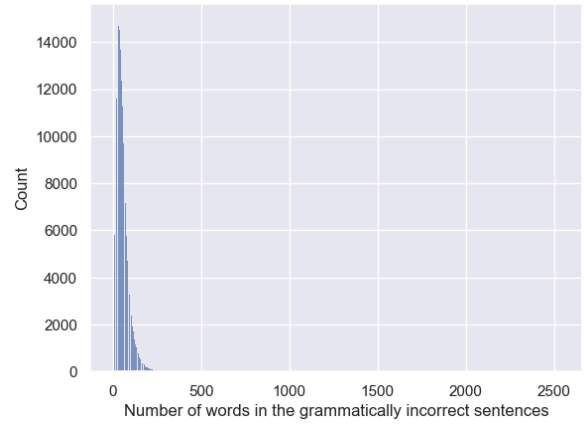


المخطط (7)

يمثل المخططان السابقان تكرار أكثر ثلاثين كلمة وردت في الجمل الصحيحة والخاطئة، تشغل علامات الترقيم المرتبة العدد الأكبر من المخططات بالإضافة لورود الكلمات المختصرة التي سبق وجودها في الـ wordcloud مثل الـ 'n't والـ 's والتي يمكن معالجتها.



المخطط (9)



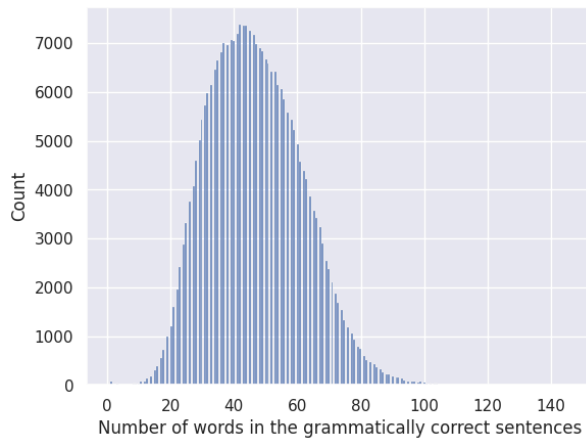
المخطط (8)

يمثل المخططان السابقان عدد العينات إلى عدد الكلمات التي تحتويها، يأخذ كلا المنحنيان شكل التوزيع الطبيعي وهذا الأمر منطقي نظرا لكون الداتا تم تجميعها من منصة لتعلم اللغة.

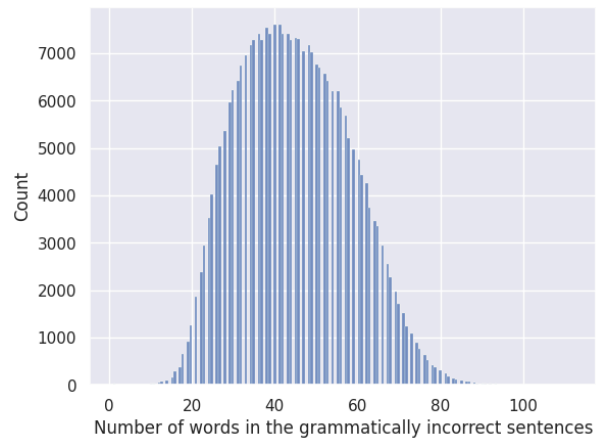
## 3.2 معالجة البيانات

اتبينا العمليات التالية عند القيام بمعالجة وتنظيف البيانات:

- استبدال الكلمات المختصرة مثل n't ب not أو 'll ب will.
- التخلص من التعابير الرسومية مثل xD وغيرها.
- إزالة الأقواس وما تحتويه ضمنها.
- التخلص من جميع المحارف التي لا تفيد والإبقاء على الحروف وبعض علامات الترقيم المهمة مثل النقطة والفاصلة وغيرها.
- التخلص من الأحرف المكررة أكثر من مرة واستبدالها بمحرف واحد.
- حذف الأرقام.
- التخلص من العينات التي يقل عدد مفرداتها عن 5 أو يزيد عن 15.



المخطط (11)



المخطط (10)

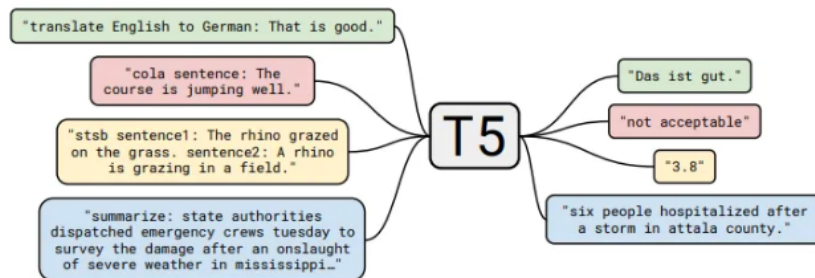
يمثل المخططان السابقان بعد تطبيق عمليات التنظيفات والتنظيف على مجموعة العينات.

## 4 تصحيح الأخطاء القواعدية باستخدام الـ transformers

لتصحيح الأخطاء القواعدية مقترح طريقتين لبناء النموذج، الأولى هي باستخدام الـ t5 transformers التي طورتها شركة google والثانية باستخدام نموذج encoder decoder and attention model، و سنتوسع بالحديث عن كلا الطريقتين في الفقرتين التاليتين.

### 1.4 نموذج T5 transformers

يُعتبر نموذج T5، المعروف بـ "نموذج تحويل النص إلى نص" (Text-to-Text Transfer Transformer)، أحد نماذج معالجة اللغات الطبيعية (NLP) التي طورتها شركة Google. يستند هذا النموذج إلى (Transformer architecture)، وهي نوع من الشبكات العصبونية التي أثبتت فعاليتها بشكل كبير في مهام معالجة اللغات الطبيعية. يتميز نموذج T5 بقابليته للتكيف بدرجة عالية، مما يجعله مناسباً لمجموعة واسعة من المهام، بما في ذلك تصنيف النصوص، الإجابة على الأسئلة، وترجمة اللغات وتصحيح الأخطاء القواعدية. نستعرض بشكل أكثر تفصيلاً نموذج T5.



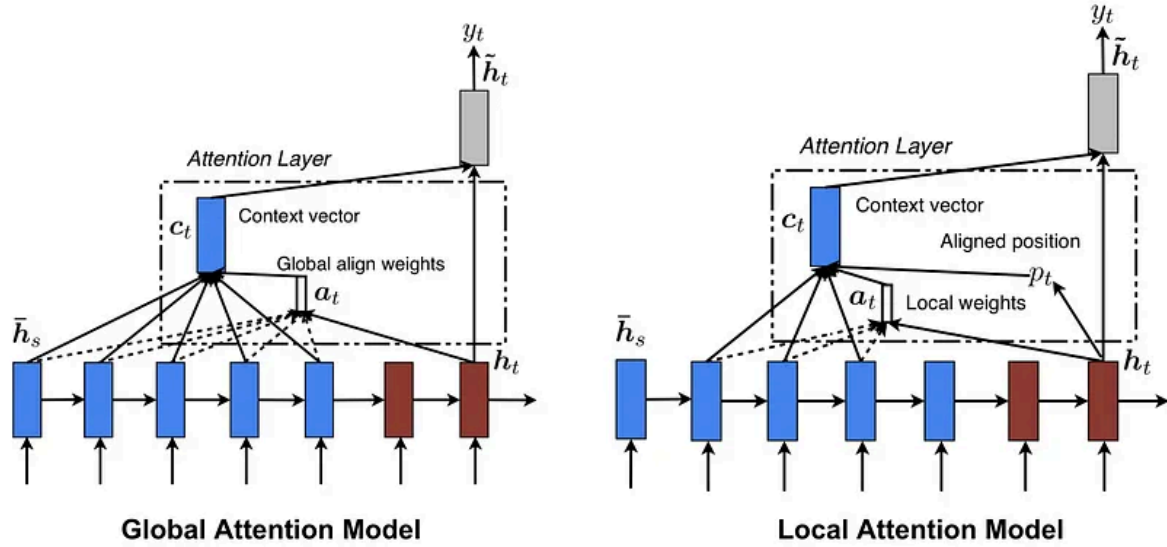
المخطط (12)

يتألف نموذج T5 من (encoder-decoder structure)، مشابهة للنماذج الأخرى التي تعمل وفق (sequence-to-sequence models). يستخدم النموذج مجموعة من (transformer layers) المؤلف من (encoder) و (decoder). يعالج (encoder) الدخل الذي يكون عبارة عن سلسلة كلمات، ويتم تطبيق (self-attention mechanisms) لالتقاط العلاقات بين الكلمات والمعلومات السياقية. تسمح هذه الآلية للنموذج بالتركيز على أجزاء مختلفة من تسلسل الكلمات الوارد إليه، مع الأخذ بعين الاعتبار الارتباطات بين هذه الكلمات من ناحية أخرى، يقوم (decoder) بتوليد تسلسل الكلمات الناتجة حيث يأخذ الخرج النهائي من الـ (encoder) كدخل، ويولد سلسلة الخرج خطوة بخطوة باستخدام آلية attention للتركيز على الأجزاء المختلفة من الدخل أثناء التنبؤ بالكلمة التالية. تستخدم كل من طبقات encoder decoder في نموذج T5 الـ (multi-head self-attention)، مما يمكن النموذج من التقاط اعتماديات مختلفة في الدخل والخرج. تعزز آلية attention قدرة النموذج على التنبؤ بخرج الكلمات حتى ولو كانت الجملة طويلة والنقاط المعلومات السياقية بفعالية. يستخدم نموذج T5 أيضاً (positional encoding) لإعادة تمثيل القيم للكلمة وموقعها في الجملة. الأمر الذي يساعد هذا النموذج على فهم ترتيب وموقع الكلمات في التسلسل، وهو أمر ضروري لمعرفة الطبيعة التتابعية للغة. أحد الجوانب المميزة لهيكلية نموذج T5 هو نهجه القائم على تحويل النص إلى نص. بدلاً من امتلاك هيكليات محددة لكل مهمة، يعامل النموذج جميع المهام في معالجة اللغات الطبيعية على أنها مشاكل تحويل نص إلى نص. هذا يعني أن كلاً من الدخل والخرج يعاملان كسلاسل نصية، مما يسمح لنموذج T5 بالتعامل مع مهام متنوعة باستخدام إطار عمل موحد.

### 2.4 نموذج Encoder Decoder and Attention Model

يتكون النموذج من جزأين أساسيين: المشفر Encoder والمفكك Decoder.

يعمل هذا النموذج بشكل أساسي على الربط بين الجمل Sequence to Sequence Mapping، حيث يقوم الـ Encoder بتحويل الـ Input Vector إلى Context Vector، ومن ثم يقوم Decoder بتحويل متجه الـ Context Vector إلى Output. على الرغم من الأداء الممتاز الذي يقدم الـ Encoder و الـ Decoder فإنه مع ازدياد طول الجمل يصعب عليه ربط الكلمات ببعضها. هنا يأتي دور الـ Attention وهنا جوه هذه الطريقة التي تعتمد على ربط كل الكلمات ببعضها للتخلص من هذه المشكلة ويوجد نوعان من الربط:



يقدم الـ Attention الأوزان في كل دورة تدريب وهذه الأوزان تخبرنا عن مقدار الأهمية التي يجب إعطاؤها لكل output [9].

## 5 النتائج و التقييم

### 1.5 معيار التقييم

لتحديد كفاءة النموذج المدرب يجب تقييم خرج هذا النموذج، معيار التقييم العام عند التعامل مع تصحيح الأغلط 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 [1]. استخدمت الطرق المرجعية في هذا العمل لأن مجموعة البيانات تتألف من جملة خاطئة و جملة صحيحة ونريد تقييم الجملة الناتجة عن النموذج استنادا إلى الجملة المرجعية الصحيحة. الطريقتان المستخدمتان للتقييم هما Rouge score و GLEU.

### Rouge score

يعتمد rouge score في التقييم على تحديد جودة الجملة الناتجة من النموذج من خلال مقارنتها بالجملة المثالية و ذلك من خلال حساب عدد الـ n-grams المتطابقة بين الجملتين [6]. يعتمد في حسابه على المعادلة التالية:

$$ROUGE - N = \frac{\sum_{S \subseteq \{ReferenceSummaries\}} \sum_{gram_n \subseteq S} count_{match}(gram_n)}{\sum_{S \subseteq \{ReferenceSummaries\}} \sum_{gram_n \subseteq S} count(gram_n)}$$

$n$  هو طول ال  $n$ -gram  
 $ReferenceSummaries$  هي مجموعة من الجمل المرجعية المقابلة للدخل (في حالتنا تكون جملة واحدة)  
 $count(gram_n)$  هي عدد ال  $n$ -grams المرادة الموجودة في الجمل المرجعية  
 $count_{match}(gram_n)$  هو عدد ال  $n$ -grams التي ترد في كل من الجمل المرجعية و الجمل التي أخرجها النموذج.

## GLEU

تحتسب GLEU الدقة الموزونة لـ  $n$ -grams من الجمل المخرجة من النموذج على الجمل المرجعية، والتي تكافئ  $n$ -grams التي تم تغييرها بشكل صحيح بينما تعاقب  $n$ -grams التي تظهر في الجملة المخرجة من النموذج ولكن ليس في الجمل المرجعية [1]. من خلال الأعمال السابقة تبين أن GLEU أكثر تشابهاً مع التقييم البشري من أساليب التقييم الأخرى [7].

## 2.5 النتائج

### نتائج الطريقة الأولى

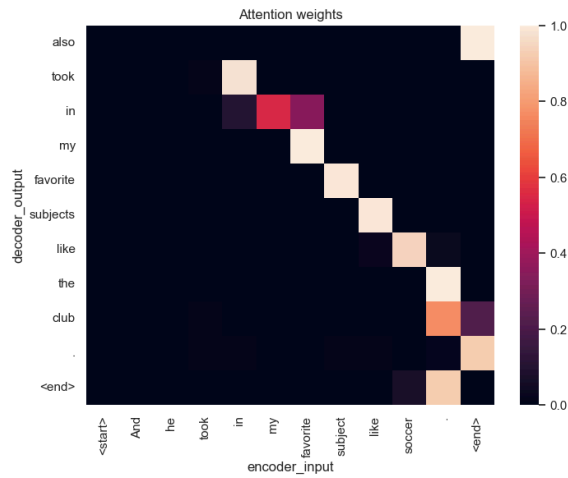
عن تدريب النموذج المعتمد على T5 باستخدام 100 ألف من عينة التدريب فقط حقق نتيجة 12.5% عند التقييم باستخدام GLEU، و 13.7% عند استخدام 170 ألف عينة من عينات التدريب. نلاحظ أن زيادة عدد عينات التدريب أدى إلى زيادة في دقة النموذج. أما بالنسبة للمقيم ROUGE حصل النموذج على دقة 83% عند التدريب باستخدام 100 ألف عينة تدريب، الدقة العالية تشير إلى تشابه الكلمات بين الدخل و الخرج و هذا الأمر الوارد حدوثه لأن جمل الخرج تختلف ببعض الكلمات فقط عن جمل الدخل (أي تختلف فقط بالكلمات المغلوطة) لذلك GLEU هو معيار تقييم أنسب لهذه المسألة.

DATA SAMPLE	EPOCHS	TIME	GLEU	ROUGE 1	ROUGE 2	ROUGE L
100k	1	23:17	12.537700	83.360600	66.805200	82.028000
170k	1	1:34:47	13.744100			

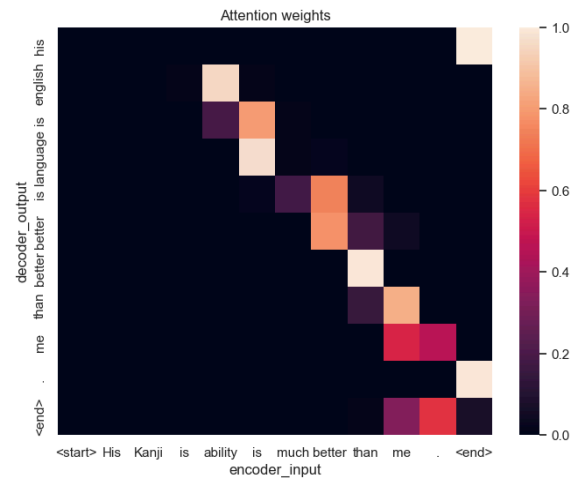
الجدول (5)

### نتائج الطريقة الثانية (Encoder Decoder model with attention):

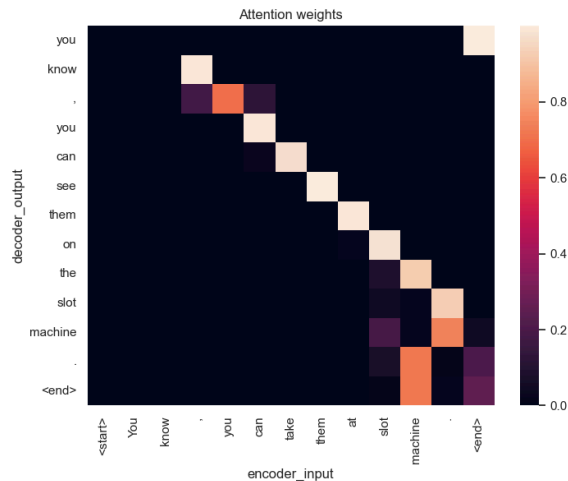
DATA SAMPLE	EPOCHS	TIME	F score
10k	10	22:16:00	0.49



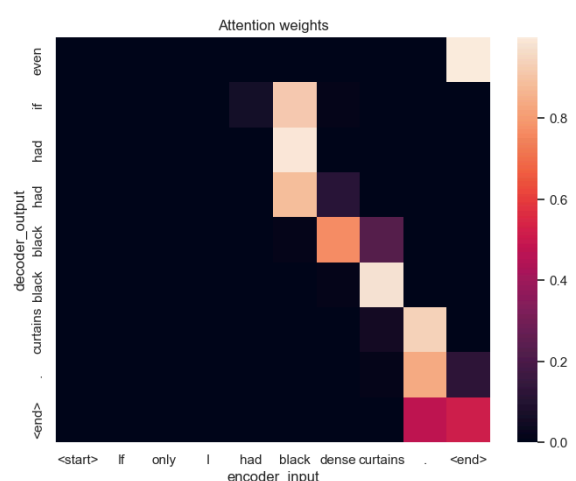
المخطط (14)



المخطط (13)



المخطط (16)



المخطط (15)

المخططات السابقة تمثل أوزان ال attention التي تم إسنادها للجملة الدخلى لكي يكون الخرج كما هو موضح فى كل مخطط و يكون ال decoder output عبارة عن سلسلة من الكلمات التي تم توليدها من قبل النموذج. ويمكن ان نلاحظ ايضا كل كلمة من الخرج كم يكون هناك مقدار attention بالنسبة لكل كلمة من الدخلى. مما سبق نلاحظ أنه باستخدام هذه الطريقة يمكننا الوصول الى نتائج أفضل ولكن الأمر يحتاج إلى وقت تدريب وموارد أكثر.

## 6 الملخص

فى هذا المشروع قمنا بمناقشة عنوان تصحيح الأخطاء القواعدية بشكل آلى. القسم الأول احتوى مقدمة عن المشكلة و هيكل العمل. فى القسم الثانى تحدثنا عن أنواع الاغلاط القواعدية و قيودها كما ناقشنا الدراسة المرجعية. القسم الثالث استعرضنا مجموعة البيانات المستخدمة من خلال EDA إضافة إلى ذكر العمليات المعالجة عليها. فى القسم الرابع قمنا بطرح طريقتين لبناء نموذج يقوم بحل المشكلة المطروحة. و فى القسم الخامس عرضت نتائج الطريقتين المذكورتين. وفقا للنتائج التي تم إيجادها عند تدريب نموذج ال T5 كان دقة النموذج الذي تم بناؤه تزداد، وهذا الأمر عائد لزيادة عدد العينات عند التدريب. أما عند بناء وتدريب النموذج الثانى فعند

تدريب النموذج على عدد قليل من البيانات نسبيا كان الزمن المستغرق في التدريب 22 ساعة. ختاماً نجد أن عملية التصحيح القواعدي للغة عملية مكلف من ناحية الموارد والوقت ولكن عند توافر هذه العوامل من الممكن تحقيق نموذج ذو دقة جيدة.

## 7 المراجع

- [1] Yu Wang, Yuelin Wang, Jie Liu, Zhuo Liu. A Comprehensive Survey of Grammar Error Correction.
- [2] Yuejiao Fei♣♣, Leyang Cui♥, Sen Yang♦, Wai Lam♦, Zhenzhong Lan♣, Shuming Shi♥. Enhancing Grammatical Error Correction Systems with Explanations.♣ Zhejiang University ♥ Tencent AI lab ♦ The Chinese University of Hong Kong ♣ School of Engineering, Westlake University.
- [3] Roman Grundkiewicz† Marcin Junczys-Dowmunt‡ Kenneth Heafield†. Neural Grammatical Error Correction Systems with Unsupervised Pre-training on Synthetic Data.† University of Edinburgh, Scotland.
- [4] D. Naber, "A rule-based style and grammar checker," 01 2003.
- [5] Courtney Napoles<sup>1</sup> and Keisuke Sakaguchi<sup>1</sup> and Matt Post<sup>2</sup> and Joel Tetreault<sup>3</sup>. Ground Truth for Grammatical Error Correction Metrics. 1 Center for Language and Speech Processing, Johns Hopkins University 2 Human Language Technology Center of Excellence, Johns Hopkins University 3Yahoo Labs
- [6] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. Information Sciences Institute University of Southern California.
- [7] Courtney Napoles, 1 Keisuke Sakaguchi, 1 and Joel Tetreault<sup>2</sup>. There's No Comparison: Reference-less Evaluation Metrics in Grammatical Error Correction. 1 Center for Language and Speech Processing, Johns Hopkins University 2 Grammarly.
- [8] Masato Mita (1,2) and Hitomi Yanaka(3,1). Do Grammatical Error Correction Models Realize Grammatical Generalization?. 1 RIKEN AIP, 2 Tohoku University, 3 The University of Tokyo.
- [9] Minh-Thang Luong Hieu Pham Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation. Computer Science Department, Stanford University, Stanford, CA 94305.