**Spelling Correction**

**وسام السحلي، عبير السيد، حلا عويس، تسنيم زغموت**

قسم الذكاء الصنعي، جامعة دمشق

تصحيح الأخطاء الإملائية في اللغة الإنكليزية هو مشروع لفت الكثير من الباحثين في معالجة اللغات الطبيعية بسبب أهميته في الكثير من التطبيقات التي نستعملها يومياً وفي هذا المشروع سنقوم ببناء مصحح تلقائي للجمل اعتماداً على السياق حيث سنقوم بتحديد الكلمة الغير صحيحة إملائياً ثم إيجاد الكلمات الصحيحة بتوليد كلمات من الكلمة المدخلة التي حددناها عن طريق معجم واختيار الكلمة الصحيحة والمناسبة حسب السياق باستخدامlanguage model سنقوم ببنائها.

**الكلمات المفتاحية :** spelling correction -contextual spelling correction-language model

**المقدمة :**

هدفنا أن نقوم ببناء نموذج يقوم بالتصحيح الإملائي اعتماداً على السياق اذ هناك أنواع مختلفة من الأخطاء الإملائية أخطاء الكلمة الحقيقية: أحيانا بدلا من إنشاء كلمة معينة ينتهي بك الأمر إلى إنشاء كلمة حقيقية ولكن كلمة لم تكن تنوي القيام بها. مثل: طباعة buckled عندما عَنيتَ bucked .أو أخطاء تنتج كلمات غير موجودة باللغة .حيث تكمن الصعوبة بهذا المشروع إيجاد الكلمة المناسبة سياقاً حيث سيكون لدي عدد كبير من الكلمات الصحيحة المولدة من الكلمة ولحل هذه المشكلة سنعتمد على language model لاختيار الكلمة المناسبة سياقاً من بين الكلمات و التحقق بحال كانت الكلمة صحيحة إملائياً إن كانت مناسبة حسب السياق.

**الأعمال السابقة :**

**1- A Benchmark Corpus of English Misspellings and a Minimally-supervised Model for Spelling Correction**

بهذا النموذج يتم التصحيح الإملائي للكلمة بحال كانت الكلمة المكتوبة غير موجودة باللغة تم استخدام (The TOFEL-SPELL Cropus) كداتاسيت بهذا المشروع، يقوم النموذج بالتصحيح اعتماداً على السياق حيث يكتشف الكلمات الغير صحيحة بنسبة 100% أما بالنسبة للتصحيح فتم استخدام أكثر من خوارزمية حيث أعطت خوارزمية ال Perceptron أعلى دقة وهي 88.12

**2- Clinical spelling correction with word and character n-gram embeddings**

clinical free-text أسلوب تصحيح إملائي حساس للسياق غير خاضع للإشراف الذي يستخدم الكلمات

والحروف المضمنة عن طريق n-gram

نولد أخطاء إملائية بديلة للمرشحين وتصنفهم وفقًا لتناسبهم الدلالي، عن طريق حساب تشابه التجيب

الموزون بين التمثيل المتجه لمرشح وسياق الخطأ الاملائي (cosine similarity)

. لقد تفوقوا بشكل كبير على أداتين أساسيتين لتصحيح الإملاء على مجموعة اختبار يدويا ومواجهة التحيز الترددي لنموذج قناة الضجيج مما يظهر أنه يمكن استغلال التضمينات العصبية بنجاح لتضمين الوعي بالسياق في نموذج التصحيح الاملائي MIMIC-III

وتقسميها إلى أربع قوائم : NOTEEVENTS from MIMIC-III v1.3 تم استخراج بيانات اختبار اللغة الإنجليزية

من لاستبدالات الصحيحة، والأخطاء الإملائية، وسياقات الأخطاء الإملائية، ومؤشرات الأسطر

بعدها تمت معالجة بيانات اللغة الإنجليزية (Preprocessing) ومن ثم تم توليد قوائم التكرارات

development corpus من البيانات المعالجة مسبقًا ومن ثم توليد وneural embedding التيتم إنشاؤها و

مرشحين لبيانات الاختبار المستخرجة أو البيانات الأخرى التي تمت ملاحظتها تجريبياً

3- **Ekphrasis**

التصحيح الإملائي : يعتمد مصحح الإملاء على المصحح الإملائي ل نستخدم إحصائيات الكلمات من أجل العثور

على المرشح الأكثر احتمالًا بحيث يتم استبدال الكلمة التي بها أخطاء إملائية بأكثر الكلمات المرشحة احتمالا

Ekphrasis هي أداة لمعالجة النصوص ، موجهة نحو النص من الشبكات الاجتماعية مثل

Twitter أو Facebook ينفذ Ekphrasis الترمي ، وnormalization

كلمات ، وتجزئة الكلمات (لتقسيم علامات التجزئة) وتصحيح الإملاء ، باستخدام إحصائيات الكلمات من

مجموعتين كبيرتين (ويكيبيديا الإنجليزية ، تويتر - 330 mil للتغريدات الإنجليزية

تعمل آليات تجزئة الكلمات والتصحيح الإملائي فوق إحصائيات الكلمات التي يتم جمعها من مجموعة معينة.

نحن نقدم حصائيات عن الكلمات من مجموعتين كبيرتين (من ويكيبيديا وتويتر) ، ولكن يمكنك أيضًا إنشاء

إحصائيات الكلمات من مجموعتك الخاصة. قد تحتاج إلى القيام بذلك إذا كنت تعمل مع نصوص خاصة بمجال

معين، مثل المستندات الطبية الحيوية. على سبيل المثال، يمكن معالجة الكلمة التي تصف تقنية أو مركبًا

كيميائيًا ككلمة بها أخطاء إملائي وذلك باستخدام إحصائيات الكلمات من مجموعة عامة مقصودة.

**البيبر التي تمت قراءتها:**

* الداتا : تم استخدام (The TOEFL-Spell Corpus) كداتاسيت بهذا المشروع حيث تتألف الداتا من المقالات المكتوبة بامتحان TOEFL iBT المستخدم دولياً. وتحوي الداتا نحو883 مقال من المقالات التي حصلت على علامة عالية أو متوسطة كما و تحتوي الداتا على 296,141 كلمة من اللغة الإنكليزية.
* نموذج التصحيح الإملائي:

تتكون مهمة التصحيح الإملائي من ثلاث مهام فرعية: كشف الخطأ، وإنشاء التصحيحات المرشحة، وترتيب المرشحين.

* كشف الخطأ: يتم الكشف عن الأخطاء الإملائية باستخدام القاموس (المعجم). الرموز التي ليست في المعجم تعتبر أخطاء إملائية. يستخدم قاموسًا يتكون من 140.000 كلمة مفردة (بما في ذلك التصريفات)، و 100.000 مصطلح متعدد الكلمات، و 130.000 اسم (بما في ذلك الأسماء والألقاب من مختلف البلدان). يتضمن القاموس كلا من التهجئة الأمريكية والبريطانية المتنوعة والاختصارات الشائعة والكلمات الأجنبية( الدخيلة ). يتضمن القاموس lexica من WordNet ، ومشروع SCOWL ، وأسماء من بيانات التعداد الأمريكية ، وقوائم Wikipedia ، ومصادر مختلفة على الويب.
* إنشاء التصحيحات المرشحة: يتم إنشاء التصحيحات المرشحة باستخدام القاموس الموضح أعلاه. تقوم التصحيحات المرشحة بتضمين جميع كلمات القاموس ضمن مسافة التعديل التي لا تتجاوز نصف طول السلسلة التي بها أخطاء إملائية ، مع مسافة أقصاها 6 أحرف. يتم إنشاء كل التصحيحات المرشحة من الرموز المميزة الفردية والمتعددة ، للسماح بتصحيح أخطاء الاندماج. لكل رمز مميز فيه أخطاء إملائية ، يتم إنشاء المئات من التصحيحات المرشحة ، باستخدام بنية بيانات شجرة البحث الثلاثية.
* ترتيب تصحيحات المرشحين : خطوة الترتيب هي الخطوة الأكثر تحديًا وهي محور معظم العمل على التصحيح الإملائي غير اللغوي. يستخدم النموذج كلاً من ميزات ( features ) الخطأ الإملائي والزوج المرشح والمعلومات السياقية. يتضمن الزوج المرشح التشابه الإملائي والتشابه الصوتي وتردد الكلمات المرشحة. تتضمن المعلومات السياقية دعم n-gram ، وتقديرًا لإعادة استخدام الكلمات في النص ، وتضمينات الكلمة.
* يتم حساب التشابه الهجائي على أنه مسافة تعديل عكسية ، 1 / (eDist + 1) ، حيث يمثل eDist مسافة التعديل (بما في ذلك عمليات التحويل) بين الخطأ الإملائي و التصحيح المرشح.
* يعكس التشابه الصوتي الحدس القائل بأن التصحيح الجيد يجب أن يكون مشابهًا صوتيًا للخطأ الإملائي. يتم حسابها على أنها 1 / (eDistP h + 1) ، حيث eDistP h هي مسافة التعديل بين التمثيل الصوتي للخطأ الإملائي والتمثيل الصوتي للتصحيح المرشح. يتم حساب التمثيلات الصوتية باستخدام خوارزمية Metaphone مزدوجة.
* تردد التصحيح المرشح يعني أنه من المرجح أن تكون الكلمة الأكثر تكرارًا هي الكلمة المقصودة من الكلمة النادرة. يتم حساب تردد كلمات Unigram لكل مرشح باستخدام مجموعة Wikipedia corpus الإنجليزية.
* دعم N-gram حيث لكل تصحيح مرشح، يتم أخذ جميع n-grams في النافذة المكونة من أربع كلمات سياق على كل جانب في الاعتبار من خلال ميزة دعم n-gram. يستخدم أعداد التكرار المحسوبة من مجموعة ويكيبيديا الإنجليزية ويتم ترجيحها على أنها درجات إيجابية لمؤشر PMI (PNPMI). تم تقديم مؤشر PMI المعياري بواسطة Bouma (2009) ، ونقوم بتعديله على النحو التالي : log2 (p(c, ngram)/p(c)p(ngram) )/(− log2 p(c, ngram)) يعين PNPMI جميع القيم السالبة إلى الصفر. لكل مرشح c ، يتم إنشاء جميع n-grams من 2 إلى 4 كلمات في إطار السياق ، ويتم إضافة قيم PNPMI لكل زوج c ، n-gram.
* ديجافو، تأخذ هذه الميزة في الاعتبار سياق المقال الواسع وتكافئ التصحيح المرشح الذي يظهر في نفس المقالة. كل تكرار للتصحيح المرشح في النص يقوي التصحيح المرشح بمقدار 1/sqrt(1 + distance) ، حيث المسافة هي عدد الرموز بين الخطأ الإملائي وموضع التصحيح المرشح في النص.
* DejavuSM هي ميزة تعمل على الأخطاء الإملائية المنهجية ، عند وجود خطأ إملائي في كلمة ما خلال المقال. لكل تصحيح مرشح ، نقوم بالبحث في قوائم التصحيحات المرشحة للرموز المميزة التي بها أخطاء إملائية أخرى في النص. في كل مرة يتم فيها العثور على التصحيح المرشح أو انعطافه في قائمة أخرى ، يتم تقوية التصحيح المرشح بدرجة SCC/sqrt(1 + distance) ، حيث SCC هي القوة الإجمالية الحالية التي تم إعادة قياسها للتصحيح المرشح المقابل في القائمة الأخرى.
* لقد أظهرت عمليات تضمين الكلمات Word embeddings نجاحًا كبيرًا في العديد من تطبيقات البرمجة اللغوية العصبية ، خاصة في تقدير الارتباط الدلالي. نحن نستخدم تضمين الكلمات لتسجيل التوافق السياقي لمرشحي التصحيح في السياق المحلي للأخطاء الإملائية. الفكرة هي أنه بالنسبة لخطأ إملائي مثل "roat" ، يجب تعزيز تصحيح " road " إذا تم العثور على كلمة مثل " drive " في المنطقة المجاورة. بالنظر إلى رمز مميز به خطأ إملائي ، فإننا نحدد نافذة من الرموز المميزة بـ 15 ± حوله. لكل تصحيح مرشح ، نحسب تشابه جيب التمام بين متجه التضمين للتصحيح المرشح ومتجه كل كلمة سياق ، ونجمع هذه القيم ، وهذه هي درجة الملاءمة السياقية القائمة على المتجه للتصحيح المرشح. نستخدم متجهات word2vec ذات 300 بُعد ، مدربة مسبقًا على 100 مليار كلمة في Google News.
* ترتيب تصحيحات المرشحين حيث لكل رمز مميز به أخطاء إملائية ، يتم تسوية درجات الميزات للتصحيحات المرشحة ، عن طريق قسمة درجة ميزة التصحيح المرشح على التصحيح المرشح الحاصل على أعلى الدرجات على تلك الميزة المحددة. يتم احتساب النتيجة النهائية لكل تصحيح مرشح كمجموع موزون لدرجات الميزة للتصحيح المرشح:

CandidateScore = P f wf · Sf حيث f نطاقات على أنواع الميزات السبعة

المستخدمة، Sf هي النتيجة المعيارية للمرشح الحالي حسب الميزة f ، و wf هو الوزن المحدد مسبقًا للميزة. يقوم النموذج الأساسي (baseline system) بتنفيذ جميع الميزات، باستثناء تضمين الكلمات. بفضل صيغة الميزة، تتطلب كل مجموعة ميزات (مثل التشابه الإملائي) وزنًا واحدًا فقط. تم اعتماد أوزان الميزات للنموذج الأساسي من Flor (2012) ، حيث تم ضبطها يدويًا. في العمل الحالي، يتم تعلم أوزان الميزات تلقائيًا باستخدام خوارزمية تعلم الآلة الخطية linear machine learning algorithm. نستخدم اثنين من المصنفات:

الخطية liner classifier

الانحدار اللوجستي Logistic Regression

والإدراك المتوسط Averaged Perceptron

**الطريقة Methodology :**

الداتا :

**تم استخدام داتا سيت** books **من** Gutenberg **فيها 20 كتاب حيث قرأنا الكتب واستخدمنا** tokenizer **من مكتبة** nltk **من نوع** punkt **لكل كتاب** **لتقسيم الكتب الى اسطر لاستخدامها في** trigram language model **وايضاً نقوم بقراءة الداتا ذاتها لكن بتقسيمها الى كلمات لاستخدامها في اختيار التصحيحات.**

**معالجة الداتا :**

قمنا ببعض عمليات الregular expirations لتنظيف الداتا ثم قسمنا كل جملة الى كلمات.

**بناء language model :**

نقوم ببناء trigram language model حيث اولاً نقوم بعمل padding لجمل الtrigram ثم نقوم بإنشاء dictionary ثم تعبئته باحتمالات ورود الكلمات بعد الكلمتين السابقتين لنقوم باستخدامه باختيار الكلمة الصحيحة لاحقاً 𝑃(𝑤3|𝑤1,𝑤2) .

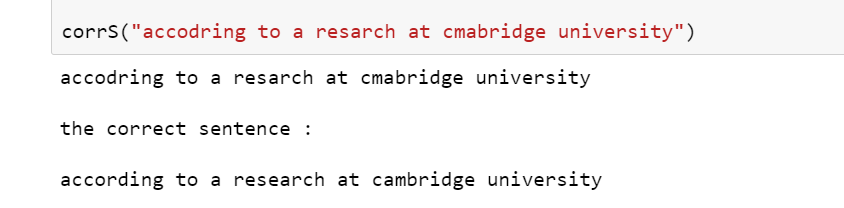
نقوم بانشاء بعض التوابع لخلق اختيارات لمصفوفة التصحيحات مثل تبديل الاحرف وإدخال احرف للكلمة وحذف احرف وثم نقوم بانشاء توابع لانشاء ثلاث مستويات من التعديلات على الكلمات.

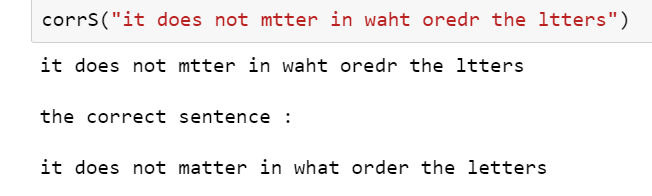
**ألية التصحيح:**

نقوم باخد الكلمات جميعها بدون التكرارات في تابع لخلق الاقتراحات فبعد اخد الاقتراحات نقوم باخد جميع الكلمات المولدة بعد الكلمتين السابقتين من ال trigram language model ثم نقوم بالمرورعلى كل اقتراح فاذا كانت موجودة بالكلمات المولدة نقوم باخد احتمالاها لمقارنتها مع باقي الاقتراحات لاخد الاقتراح الأعلى احتمالاً وجعله الخرج الصحيح

لكن اذا لم يوجد أي كلمات مولدة من ال trigram language model نقوم باخد الكلمات جميعها بدون التكرارات ثم نحسب عدد تكرارات كل كلمة واحتمال ورودها بقيامنا بتقسيم عدد ورود الكلمة على عدد الكلمات في الداتا كلها لاستخدامها في تابع لخلق الاقتراحات واحتمال كل اقتراح فبعد اخد الاقتراحات واحتمالاتها نقوم باختيار الكلمة التي احتمالاها اعلى من الاقتراحات لتكون الكلمة الصحيحة

**النتائج :**





**الخلاصة** :

قمنا ببناء نموذج يقوم بتصحيح الأخطاء الإملائية اعتماداً على السياق سواء كانت كلمة صحيحة إملائياً وسياقها خاطئ أو كلمة خاطئة وغير موجودة باللغة ونسعى بالمستقبل لبناء نموذج يقوم بتبديل الجملة الركيكة بجملة أقوى لغوياً حيث ستفيد هذه الميزة المتحدثين الغير أصليين باللغة الإنكليزية مما يساعد على تقوية لغتهم.

**المراجع** :

[**https://web.stanford.edu/~jurafsky/slp3/B.pdf**](https://web.stanford.edu/~jurafsky/slp3/B.pdf)

[**https://aclanthology.org/W19-4407.pdf**](https://aclanthology.org/W19-4407.pdf)