تحسين خوارزميّة التّجذِّيع الخفيف ArabicStemmer

ربيحة كمشة، لخضر بن زاهيّة، سحيل هشام، عاصم شلي.

تاريخ الإرسال:02 أكتوبر2021 تاريخ القبول:08 أكتوبر 2021

الملخّص Abstract: يواجه تجذيع اللغة العربيّة في مجال معالجة اللغات الطّبيعيّة تحديّات عديدة منها صعوبة تجذيع جمع الكسور وجمع المؤنث السّالم والكلمات التي في آخرها تنوين وغيرها. ومن هنا، جاء هذا العمل بهدف تسليط الضّوء على هذه الإشكاليّات ومحاولة معالجتها للتّغلب على الصّعوبات المذكورة.

وقد استعملنا، لتحقيق هذا الهدف، دراسة خصائص اللغة العربيّة وخوارزميّات تجذيعها كما فصَّلنا مراحل المجذّع ArabicStemmer والذّي يمثّل أساس هذا العمل لاستخراج نقائصه ومعالجتها، مستعينين في ذلك بمدوّنات اللغة العربيّة ودراسة مراحل المعالجة المسبقة للنّصوص.

وقد أظهرت هذه الدّراسة حدوث عدّة تحسينات، أهمها:

- أن التّعديل الذّي احدثناه على (ArabicStemmer) أدّى إلى تحسين أداء التّجذيع بنسبة 4.2 % مقارنة بالنّسخة الأولى.
- كما توصّلنا إلى استخراج 7232 جذعاً صحيحاً من بين 18350 كلمة في المدوّنة القرآنيّة، أي ما يقابل 39.41 % على الرّغم من استعمال المدوّنة للأشكال القديمة للحروف العربيّة مع التّشكيل.
- وفيما يخص اللغة المعاصرة، تمكنا من إيجاد 66.72 % جذعاً صحيحاً من كلمات المدوّنة الذّهبيّة مقابل 27.60% لثاني أحسن مجذّع المعروف بـ"تاشفين" للباحث طه زروقي.

وكخلاصة، توصلنا إلى أن عملنا هذا أحدث تحسناً في أداء الخوارزميّة في تجذيع الأوزان البسيطة، والمدوّنة القرآنيّة والمدوّنة الذّهبيّة.

المدخل Introduction: اللغة العربيّة من أكثر اللغات العالميّة غِنَى من حيث مورفولوجيتها، "*وتقدر ذخيرتها اللغويّة بأكثر من اثني عشر (12) مليون كلمة*" (روينة, 2019) يتحدث بها حوالي مليار ونصف شخص في جميع أنحاء العالم، وتعتبر واحدة من اللغات الرّسميّة السّت للأمم المتحدة. وعلى الرّغم من هذه الأهميّة البالغة إلا أنها حظيت باهتمام ضئيل نسبيا في مجال معالجة اللغات الطّبيعيّة (Natural Processing Language)، وهو المجال الذّي يسمح للحاسوب بتحليل وفهم واستخلاص المعاني من لغة الإنسان الطّبيعيّة باستعمال تقانات مرتكزة أساساً على علم اللسانيّات والذّكاء الاصطناعي. ومن أبرز العمليّات الأوليّة التي تتم في هذا المجال ما يعرف بالمعالجة المسبقة للنّصوص حيث ظهر هذا المفهوم لوجود كمٍ

هائلٍ من النّصـــوص في العالم الرّقمي، والتي يستهدفها المستخدمون إمّا لتعديلها أو تحليلها أو غير ذلك .

تهدف المعالجة المسبقة للنّصوص إلى تهيئة النّص للمعالجة والتّحليل وضمان جودة العمليّات المزمع اجراؤها على النّص كالفهرسة واسترجاع المعلومات وضغط المستندات وتحليل النّصوص الإحصائيّة ومجالات أخرى خاصة بالمعالجة الآليّة للغة الطّبيعيّة.

تمر المعالجة المسبقة للنّصوص (زلاقي, 2018)بعدّة مراحل كالفصل إلى كلمات حسب الفراغات الموجودة، أو علامات الوقف (Chelli, 2012)، وتصفيّة كلمات الوقف بحذف علامات الوقف (الفاصلة، الفاصلة المنقوطة، علامة الاستفهام، حروف الجر، الضّمائر، والأسماء الموصولة...) حتى لا يتم فهرسة كلمات الوقف في ملف نصي أثناء تخزين واسترجاع البيانات . هناك أيضاً مرحلة التّطبيع التي تعد عمليّة أساسيّة في المرحلة الأوليّة لمعالجة اللغة الطّبيعيّة، بحيث تتبع عدّة معايير تُحَدَدُ من طرف خبراء اللغة، من بينها : 1.حذف علامات التّرقيم 2.إرجاع الحروف إلى أصلها المنفصل .ففي اللغة العربيّة ، تَرِدُ الحروف في عدّة صور، وذلك حسب موقعها في الكلمة .**مثال** : صور الحرف' ع 'هي :ع، ـعـ ، عـ ، ـع، كل هذه الصّور عند التّطبيع تصبح ع.

مرحلة أخرى من مراحل المعالجة المسبقة للنّصوص مرحلة **إيجاد المدخل المعجمي** : حيث تهدف هذه العمليّة إلى إيجاد الأشكال المختلفة للكلمة المشتقّة كما هي في قواميس اللغة، ويطلق على هذه الأشكال" ليما(Manning *et al.*, 2010) (Lemma) " مثال: الكتابة نجدها في القاموس اللغوي : كِتَابَة ، كِتَاب...

أما آخر مرحلة وهي التّجذيع وتتمثل في كشف الحروف الأصليّة وإزالة جزء من السّوابق واللواحق وبعض الحروف المحشوة حتى يسهل البحث عن الكلمات الممثلة بنفس الجذع.

1.1. خصائص اللُّغة العربيّة: تعتبر الكتابة العربيّة من اللغات الأكثر غنًى حيث تضمّ "*أكثر من إثني عشر(12) مليون كلمة بالإضافة إلى الاشتقاق والنّحت والقياس والتّوليد والتّرجمة مما يجعلها قادرة على استيعاب مختلف العلوم والتّقانات ومُسايرتها*"(روينة, 2019).

وتعتبر كتابتها كتابةً شبه مخطوطة، إذ أن معظم الحروف ملتصقة مع بعضها البعض عدا الهمزة 'ء' كما توجد ستة حروف فقط من أصل 29 لا تلتصق بها الحروف التي تليها وهي 'و ز ر ذ د ا '. والجدير بالذّكر هنا، كما أشار إليه الأستاذ الدّكتور روينة، أن حروف العربيّة لا تتعدى 16 حرفاً إذا ما أزلنا علامات التّنقيط، فالباء والتّاء والثاء تكتب باستعمال نفس الشّكل 'ـب' وتختلف في عدد وموقع النّقاط ليتغير بذلك شكلها وصوته ولإزالة الغموض في نطق الكلمات العربيّة توضع واحدة من علامات التّشكيل: الكسرة، الضّمة أو الفتحة اختياريا بعد كل حرف، ويستعمل السّكون في حالة عدم وجود إحدى هذه العلامات. أما عند وجود حرفين متشابهين تكتب الشّدة فوق أحدهما ويحذف الآخر. كما أنه يجدر الإشارة إلى أن الكلمات في اللغة العربيّة تنقسم إلى 3 أقسام وهي الاسم والفعل والحرف.

وهناك عدّة قواعد لإعراب الأسماء، كإعراب الأدوات وإعراب الأسماء المفردة، وأسماء المثنى وأخيراً أسماء الجموع. توجد أسماء الجموع على صنفين هما:

- **الجموع السّالمة** (بن جنّي, 1988) : تنشأ من خلال إضافة لاحقة إلى الصّيغة المفردة دون التّغيير في الكلمة.
  - **جمع المذكر السّالم** :نضيف "ون" في حالة الرّفع، و"ين" في حالة النّصب والجر .مثال : جمع" مدرس "يصبح " مدرسون " أو" مدرسين."
  - **جمع المؤنث السّالم** :نضيف الحرفين 'ات 'للحصول على الجمع .مثال :جمع" سيارة " يصبح" سيارات."
- **جموع التّكسير**: (الأنباري, 1995) هو ما يدلّ على أكثر من اثنين وتغيّرت فيه صيغة المفرد، ويكون هذا التّغيير اعتمادا على الاسم، ممّا يؤدّي إلى وجود قواعد مختلفة لجموع التّكسير .مثال :جمع الاسم " كاتب "يكون " كتّاب "أو " كتبة ."كما أن الجموع التي تعتمد على الأعداد أيضا تعتبر جموع التّكسير، مثل : جمع" شهر "هو " أشهر "أو" شهور."

كل هذا التّعدّد في القواعد يشكّل على حدّ سواء غنى اللغة العربيّة وحاجزاً في انجاح البرمجيّات المصمّمة للمعالجة الآليّة اللغة. إضافة إلى الاختلافات الإقليمي الإملائيّة وعدم توحيد التّقنيّات المستعملة لحوسبتها. ومن أهمّ المعوّقات لتوفير الآليّات والتّقانات لمعالجة اللغة العربيّة هي قلّة المتخصّصين في اللسانيّات العربيّة وعلم الحاسب في آنٍ واحدٍ ممّا يزيد في الهوة المتواجدة بين هذين العلمين. ومن أجل هذا وللارتقاء بالعربيّة، أتت بحوث من طرف باحثين عرب واجانب للنّهوض بحوسبة العربيّة وإقحامها ضمن برامج المعالجة الآليّة للغة.

ونهدف، في الفقرة .2.I. المواليّة، إلى إماطة اللثام عن مرحلة مهمّة من مراحل المعالجة المسبقة للنّصوص ألا وهي التّجذيع والذّي يعتبر لبّ بحثنا هذا.

2.I. التّجذيع Stemming: تَستعمل الملفات النّصيّة أشكالاً مختلفة للكلمات، ويعود ذلك لأسباب نحويّة، فمثلا نجد في نصّ واحد عدّة أشكال لكلمة "نَظَمَ" كالتّالي (منظّم/ يُنظّم /تنظيم/ منظّمة)، ويطلق عليها عادة مصطلح العائلة .وعليه، عائلة الكلمات هي مجموعة من الكلمات المشتقّة من كلمة واحدة تدعى الأصل وتُعْرَفُ عند اللغويين بالجذر .أمّا بالنّسبة للمعالجة الرّقميّة للّغة الطّبيعيّة فيوجد الجذر والجذع، حيث لا يشترط في الجذع أن يكون صحيحا لغويا. ويفيدنا الجذع أو الجذر في عمليّة البحث لاسترجاع هذه الكلمات المشتقّة بدلا من البحث عن كل كلمة مشتقّة على حدة . مثال:

- كلمة مدرّس هي الجذع الذّي يمثّل العائلة :المُدَرِسُون، المُدَرِسَان، كَالْمُدَرِس، والمدرسة...
- كلمة درس هي الجذر الذّي يمثل العائلة :مدرس، المُدَرِسُون، المُدَرِسَان، كَالْمُدَرِس، والمدرسة.

وعرف (Manning) (Manning *et al.*, 2010) التّجذيع على أنّه عمليّة تقليص الكلمات المشتقّة إلى كلمة تمثّل عائلة من الكلمات، وان

لم تكن صحيحة لغويّا ، غالبا ما تكون عمليّة التّقليص بحذف السّوابق واللواحق والحروف المحشوة في الكلمات المشتقّة ؛ الحروف غير الأصليّة في بدايّة الكلمة تسمّى السّوابق، والتي في نهايتها تسمى اللواحق، أما غير الأصلي في الوسط فَيسمى الحروف المحشوة.

يساعد التّجذيع، في تحليل النّصوص الإحصائيّة، على رسم التّغييرات النّحويّة للكلمة . مثال :بعد تَجذيع "المدرسون" نستنتج من وجود الواو والنّون أنَّه جمع مذكر سالم .تستعمل المجذّعات أيضاً للتّقليل من حجم الوثائق، إذ يمكننا تخزين كلمات كبيرة في شكل جذوعها، وعند الحاجة إليها يكفي استعمال برنامج لإعادة المستند إلى شكله الأصلي، ويتم هذا عن طريق استخدام قواعد لتحديد الشّكل الأصلي للكلمة. كما يعتبر التّجذيع مرحلة أساسيّة في المعالجة المسبقة للنّصوص فعندما يتعامل نظام بحث المعلومات، مثلاً، مع الكلمات بدون استخدام مجذّع أو باستخدام مجذّع غير فعال، فان نفس الكلمات الدّالة تنتج الكثير من المتغيرات، التي بدورها تستهلك حجماً ووقتاً كبيرين مما يؤدي إلى رداءة نتائج البحث.

وعليه، تجابه المجذّعات تحديّات مختلفة في معالجة اللغات بسبب تعقيدها وتعدد قواعدها. هذا من جهة، ومن جهة أخرى، تتّسم اللغة العربيّة بمورفولوجيا ذكيّة مقارنة باللغة الانكليزيّة، على سبيل المثال للتّعبير عن الملكيّة في اللغة العربيّة، يضاف الحرف' ي 'كَلاحقة للكلمة، حيث لا توجد كلمة منفصلة توافق ' my' في الانكليزيّة. كما لا تخضع كل الكلمات إلى نفس القواعد، مثل جموع التّكسير، والمئات من الجذور التي بتغيرها واضافة حروف إليها تنتج آلاف الكلمات المشتقّة؛ وعليه، تواجه المجذّعات العربيّة على عكس الانجليزيّة أو الفرنسيّة صعوبات كثيرة. ونعرض فيما يلي أهم خصائص لغة الضّاد وأبرز التّحديّات التي تواجهها.

1.2.I. خوارزميّات عامة للتّجذيع

Stemmers: تُعْرف الخوارزميّات التي تقوم بهذه العمليّة بخوارزميّات التّجذيع أو المجذّعات وقد اقترح عدّة باحثين مجذّعات مختلفة للغة العربيّة، وكان لكل منهم طريقته في كتابة هذه المجذّعات، حيث كان الاختلاف بينها اما في المنهجيّة المتبعة في كتابة المجذّع، او في لغة البرمجة المكتوبة بها، أو حتى في المجال الذّي سيستعمل فيه هذا المجذّع .تمت كتابة عدّة مجذّعات عامّة لعدّة لغات مختلفة، وفيما يلي أشهرها:

اقترح **(Lovins)** (Lovins, 1968) مجذّع للكلمات الانكليزيّة سنة 1968 . يقوم بحذف أطول لاحقة مرتبطة بالكلمة باستخدام 252 لاحقة متغيّرة. يستطيع المجذّع التّعامل مع الكلمات التي تحتوي على حرفين متكَررين، مثل كلمة " getting ":يحولها إلى"get". إلّا أنّه لا يستطيع تغطيّة الكثير من اللواحق.

في سنة 1974 قام)(Dawson, 1974 **(Dawson)**باقتراح مجذّع يعتمد على مجذّع **(Lovins)** إلا أنه يستعمل قائمة قواعد للواحق تحتوي تقريبا 1222 لاحقة مخزنة عكس ما تم تأشيرها عليه(مؤشرة بآخر حرف وطول اللاحقة تَفَوق هذا المجذّع على مجذّع (Lovins) من حيث تغطيته لعدد أكبر من اللواحق وأيضا من حيث سرعة تنفيذه.

يعتبر (مجـذّع Porter) (Porter, 1980) من أشهر المجذّعات التي أُقتُرِحت سنة 1980 . يرتكز المجـذّع على فكرة أســاســيّـة، هي أن معظم اللواحق في اللغة الانكليزيّة تتركب من مجموعة لواحق بسـيطة .يقوم باتّباع عدّة خطوات، لكلّ منها قواعد خاصّــة، وعند تحقّق القاعدّة يتمّ حذف اللاحقة. مجموع القواعد التي اعتمدها (Porter)في مجذّعه هو 92 قاعدّة.

2.2.I. مـجـذّعـات الـلـغـة الـعـربـيّـة Arabic stemmers: توجد عدّة منهجيّات لتجذيع اللغة العربيّـة, من بينهـا: التّجـذيع الخفيف, التّجـذّيع القائم على الجذور, وغيرهما. التّجذيع الخفيف هو عمليّة إزالة مجموعة صـغيرة من السّــوابق واللواحق، بـدون التّعـامـل مع وســط الكلمـة، وكذلك دون التّعرف على الأوزان و إيجاد الجذور. أمـا التّجـذيع المعتمـد على الجـذور فيســتعمـل التّحليل الصّرفي من أجل استخراج الجذر الذّي يعطي الكلمة العربيّة.

قام (Larkey)وفريقه (Larkey *et al.*, 2007) بإنشــاء عدّة مجذّعات خفيفة لِلغة العربيّة، تقوم بحذف مجموعة من السّــوابق واللواحق، وقـاموا بتقييم فـاعليتهـا في السّــنوات 2001 و 2002. واعتمـدوا في بعض تجـاربهم على **تطبيع** الكلمـات بحـذف علامـات التّرقيم والتّشــكيـل وإزالــة الرّموز (كـل مـا يعتبر ليس من الحروف)واســتـبـدال بـعـض الحـروف بـحـروف أخـرى كاســتـبـدال الحرف الأخير 'ة' ب 'ه' وحـذف اللواحق هـا، إن، ات، ون، ين، يـه، يّـة، هـ ، ة، ي، بشـرط أن يبقى طول الكلمة حرفين أو أكثر . لم يعالج هذا المجذّع مشكلة جموع التّكسير.

أُطلِق مجذّع **(ISRI)** (Taghva *et al.*, 2005) من طرف : كاظم تـاغفـا وفريقـه بمعهـد **(ISRI)** جـامعـة نيفـادا بـالولايـات المتحـدة الأمريكيّـة. يشــترك مجذّع **(ISRI)** مع مجذّع خوجة في كثير من الخصــائص، إلا أنـه يختلف عنـه في أمر رئيسي ألا وهو عدم استعماله للقوائم.

قـام بـتـنـفـيـذ مـجـذّع (Zerrouki, 2020) **(Tashaphyne)**الباحث الجزائري طه زروقي في ســنة 2010 ، يقوم بالتّجذيع الخفيف واعطاء ابســط جذع ممكن. اســتعمل من طرف عدّة أعـمـال عـربـيّـة مـثـل :مـنـصــة (SAFAR) و (PyArabic). حيـث حَـازت هـذه الأخيرة عـام 2016على جائزة الشّـيخ محمد بن راشـد للغة العربيّة في فئة المحتوى الرّقمي.

يعمل المجذّع على تطبيع الكلمة ثم تقطيعها إلى جميع الحالات الممكنة لاســتخلاص الجذور والجذوع. يســتعمـل قـائمـة مســبقـة للزوائـد وحروف الزيادة مع إمكانيّة تبديل القائمة قبل بدايّة التّجذيع حيث يصــبح قادراً على التّجذيع حسب القائمة المزود بها لكنه لا يراعي التّشكيل.

يســتخـدم مجذّع البـاحثـة شــيرين خوجـة (Khoja, 2003)بعض خصـائص اللغة العربيّة لِاستخراج الجذور، وفق الخطوات التّاليّة:

- حـذف علامـات التّشــكيـل، كلمـات الوقف، التّرقيم والأرقام؛
- حذف أداة التّعريف' ال' و حذف حرف العطف " و"؛
- حذف اللواحق والسّوابق؛
- مطابقة النّتيجة مع قائمة الأوزان، إذا وجد تطابق تســتخرج من الوزن الحروف التي

تطابق الجذر ومقارنته مع قائمة جذور صحيحة معدّة مسبقا.

أُنشِئ مجذّع **(ArabicStemmer)** (Chelli, 2012) من طرف الباحث الجزائري شلي عاصم في مشروع Summer of Code Google قصد دعم وتحسين أداء خوارزميّة الاسترجاع XAPIANللغة العربيّة.

من خلال دراستنا للمجذّعات المعروضة، يتّضح لنا أنها لم تتعامل مع الأفعال المُعتلَّة وجُموع التّكسير (جزئياً) رغم أنها كثيرة الاستعمال في اللغة العربيّة كما لاحظنا وجود عدّة تداخلات عند حذف بعض الحروف المضافة في الأوزان، حيث قد تكون حروف أصليّة مثل :تجذيع كلمة" مائل "يعطي" ائل."

الهدف من هذا البحث هو الإسهام في معالجة التّحديّات التي تجابه جودة تجذيع اللغة العربيّة وخاصةً:

1- جمع التّكسير.
2- جمع المؤنث السّالم.
3- التّشكيل: الكلمات التي تنتهي بتنوين.

ولتحقيق هذه الأهداف، اجرينا بحثنا هذا باستعمال المواد والطّرائق الموضّحة في العنوان الموالي.

المـوادّ والـطّـرائـق Materials and Methods

لإجراء تحسينات على مجذّع اللغة العربيّة (ArabicStemmer)، اتّبعنا الخطّة التّاليّة:

1. استخراج مراحل خوارزميّة (ArabicStemmer)التي تعاني من نقص في جودة التّجذيع.
2. إحداث تعديلات في المراحل المستخرجة.
3. تنفيذ الخوارزميّة المعدلة.
4. مقارنة مخرجاتها مع مخرجات الخوارزميّة الأصليّة والمجذّعات العربيّة.

II .1. مراحل المجذّع ArabicStemmer ونقائصه

يتبع مجذّع (ArabicStemmer[1]) منهجيّة التّجذيع الخفيف، أي أنه يقوم بالتّعامل فقط مع السّوابق واللواحق، حيث يتم حذفها بعد تحديد نوع الكلمة : فعل، اسم وحالة التّعريف لاتخاذ المراحل اللازمة لتجذيع الكلمة .وهو سلوك فعال وله فوائد، خاصة إذا كان سيتم إعادة استعمال المجذّع في إنشاء محللات صرفيّة لمعرفة نوع الكلمة .

يعتبر مجذّع الباحث شلي المجذّع الرّسمي والمعتمد للّغة العربيّة في مجموعة مجذّعات إطار عمل كرة الثلج (Snowball)[2] (Porter, 2001) وهي منصة من تطوير (Porter Martin) وفريقه لتجذيع لغات عديدة. كما أن أكبر مكتبة لاسترجاع المعلومات (Xapian)[3] تعتمد عليه في معالجة اللغة العربيّة.

قام الباحث بتوفير شيفرته بعدّة لغات برمجة (c++، java، jsx، python) مع إمكانيّة تصريفه إلى عدّة لغات برمجة .يخضع التّجذيع باستعمال خوارزميّة (ArabicStemmer) إلى الخطوات الموضّحة الخوارزميّة الموالية:

[1] arabicstemmer.com
[2] snowballstem.org
[3] xapian.org

الخوارزمية: Arabic-stemmer Light Stemming Algorithm.
الهدف: تجذيع كلمة عربية.
المدخل: كلمة عربية.
المخرج: أحسن جذع ممكن.
الإجراء:
1. الترميز باستعمال Unicode ونظام تهجئة.
2. التخمين على نوع الكلمة (اسم/فعل) وحالة التعريف.
3. تطبيع أولي للكلمة.
4. حذف اللواحق إن وجدت، وذلك حسب نوع الكلمة.
5. حذف السوابق إن وجدت، وذلك حسب نوع الكلمة.
6. تطبيع نهائي للكلمة.
7. إرجاع أحسن جذع ممكن.

خوارزميّة 1 .الطّبعة الأولى لـ ArabicStemmer

II .1 .1.مراحل المجذّع:

– المرحلة 1 : يستعمل هذا المجذّع نظام التّرميز (Unicode[4]) للأحرف العربيّة كما يستعمل نظام تهجئة مقترح من صاحب المجذّع نفسه .حيث أنّ عمليّة تهجئة للحروف العربيّة باستعمال الحروف اللاتينيّة قصد تسهيل تعامل الآلة مع الكلمات العربيّة؛

– المرحلة 2 : تحديد نوع الكلمة (فعل أو اسم) ؛

- طول الكلمة >4 وتبدأ ب 'بال' و 'كال 'تعالج كأسماء معرفة؛
- طول الكلمة >3 وتبدأ ب' لل' و' ال 'تعالج كأسماء معرفة؛
- طول الكلمة >2 وتنتهي بتاء مربوطة تعالج كأسماء؛

– المرحلتان 3 و 6 : نلاحظ أن التّطبيع في هذه الخوارزميّة مقسم إلى تطبيع أولي وتطبيع نهائي فالتّطبيع الأولي يعتمد على حذف علامات التّشكيل والتّرقيم أما النّهائي فيعنى بتطبيع كل أنواع الهمزات في أول، وسط وآخر الكلمات؛

– المرحلتان 4 و5: تتعلقان بالتّجذيع .حيث قام الباحث بتقسيمه إلى نوعين :مرحلة عكسيّة

[4] نظام ترميز عالمي للنّصوص ك UTF-8 و (unicode.org) UTF-16

لمعالجة اللواحق ومرحلة مباشرة لمعالجة السّوابق. يقوم المجذّع باختيار المرحلة حسب نوع الكلمة :فعل أو اسم حسب حالة التّعريف، ويتم تحديد ذلك في مرحلتين تقومان بالتّخمين على نوع الكلمة.

– المرحلة 7: بعد المعالجة يقوم المجذّع بإرجاع أحسن جذع ممكن.

– ملاحظة: بالنّسبة لِعَمليتي **الفصل إلى كلمات واستبعاد كلمات الوقف** هما مفصولتان عن هذه الخوارزميّة، تُنفذا قبل عمليّة التّجذيع.

II .2. 1. نقائص المجذّع : والمقصود بها هو النّقائص التي استخرجناها من مراحل المجذّع وهي:

– عدم تجذيع بعض الكلمات بطريقة صحيحة مثل :مثنى المؤنث؛

– لم يعالج هذا المجذّع جمع التّكسير، بالإضافة إلى الكلمات المكونة من حرفين؛

– عدم معالجة الأوزان؛

– عدم معالجة جموع التّكسير وحذف الحروف المحشوة مثل (نمور، كواسر، شموس)؛

– عدم معالجة بعض الأسماء على الأوزان التّاليّة (اسم الفاعل والمفعول، اسم الآلة مثل فعالة... الخ)؛

– عدم دراسة أوزان المؤنث مثل :فعلة، فاعلة، فعلاء، فعلى، فعالة... الخ؛

– عدم معالجة الكلمات التي تحتوي على التّنوين كإسم؛

– عدم معالجة الإدغام وعدم معالجة الكلمات المكونة من حرفين؛

– مشاكل السّوابق مثل الخلط في بعض الحالات بين الفاء الزائدة في السّوابق والفاء الأصليّة وكذلك نفس المشكلة بالنّسبة للواو والكاف؛

– عدم معالجة الكلمات حسب الجنس مذكر/مؤنث؛

– عدم معالجة الجذور المعتلة؛

– عدم معالجة الفعل المهموز.

II .2. التّعديلات المقترحة : بعدما تعرفنا على قواعد اللغة العربيّة، والمجذّعات العربيّة التي تم تنفيذها من طرف الباحثين في مجال NLP، وتحليلنا للعمل السّابق للمجذّع (ArabicStemmer)، توصلنا إلى عدّة نقاط ستحسن أداء هذا المجذّع، خاصة في تعامله مع جموع التّكسير واسم الفاعل والمفعول...الخ، وكذلك مع الأفعال المعتلة.

II .1.2. الانتقال من التّجذيع الخفيف إلى التّجذيع القائم على الجذور: اعتمد أغلب الباحثين وعلى رأسهم شيرين خوجة على منهجيّة التّجذيع المعتمد على جذور في حل مشكلة استخراج جذور من جموع التّكسير والأسماء والأفعال التي لها أوزان معروفة في اللغة العربيّة، والتي تستعمل الأوزان في استخراج الجذور .وهنا ايضا وجد اختلاف في الطّريقة التي يعتمدونها في تطابق الأوزان، فهناك من استخدم قوائم مع التّخمين عليها مثل مجذّع خوجة، وهناك من اكتفى بالتّخمين ولم يستعمل قوائم وأشهرها هو (ISRI). يوضح

الشّكل الموالي مثال عن عمليّة التّجذيع المعتمد على الجذور.

	اللاحقة		الجذع				السابقة	
الكلمة	ت	ا	ب	ل	ا	ط	ل	ا
الوزن	x	x	ل	ع	ا	ف	x	x
الجذر	x	x	ب	ل	x	ط	x	x

شكل 1 .عمليّة التّجذيع المعتمد على الجذور

سنتطرق فيما يلي إلى الأوزان التي يمكن معالجتها لكي يصبح المجذّع له القدرة على استخراج الجذور ألا وهي **الأوزان البسيطة** . **الأوزان البسيطة** (Taghva et al., 2005) هي أوزان ذات الطّول 4 (مفعل، فاعل، ...) و 5(افتعل، مفعول، ...) و6 (مستفعل، مفعالة، ...) ونستخرج منها جذوراً بطول 3 أحرف على وزن "فعل".

II .2.2.التّخمين على نوع الكلمات : جعل المجذّع يقوم بالتّخمين والتّجذيع حسب احدى الحالات التّاليّة:

II .1.2.2. الاسم:

« فالاسم يعرف بالخفض والتّنوين ، ودخول الألف واللام، وحروف الخفض، وهي: من، والى، وعن، وعلى، وفي، وَ رب ، والباء، والكاف، واللام، وحروف القسم وهي :الواو، والباء، والتّاء» (بن آجروم, 2002)

أ. **التّنوين:** كما يعرف عند اللغويين أن التّنوين من علامات الاسم .أيضا لا يمكن ان يكون معرف ب" ال " التّعريف وفي نفس الوقت يحتوي على التّنوين.

ب. **الخفض:** يعتبر الخفض من علامات الاسم فإذا وجد الخفض في آخر الكلمة سنتعامل معها على أنها اسم.

ج. **مثنى المؤنث:** هو كل اسم دل على شخصين أو شيئين مثال :فلاحين، فلاحتين.

د. **جمع المؤنث السّالم:** هو كل اسم دل على ثلاثة أشخاص، أشياء، بزيادة ألف وتاء في آخره، مثل :طالبات، فاطمات، الجزائريّات.

II .2.2.2. الفعل: « *والفعل يعرف بقد، والسّين، وسَوفَ، وتاء التّأنيث السّاكنة*» (بن آجروم, 2002)

أ. **السّين:** والمقصود بها السّين الدّالة على الاستقبال مثل :سيفرح المجتَهد عندَ الامتحَان . فكلمة " يَفرَح " فعل، وعلامة الفعل في هذه الكلمة هي السّين.

ب. **تاءُ التّأنيثِ السّاكنة:** وهي تاء ساكنة تتصل بآخر الفعل مثل :فازت مريم بالجائزة .

فاتصـال هذه التّاء ب" فاز " يدل على أن هذه الكلمة فعل.

II .3.2. **الجمع** : الجمع يـدل على أكثر من اثنين وهو ثلاثة أنواع:

أ. **جمع مذكر سالم** :مسلمون، مسلمين

ب. **جمع مؤنث سـالم** :مسـلمات .زهرات، استاذات، طالبات.

ج. **جمع تكسير** :هو الجمع الذّي تتغيّر فيه صـورة المفرد بزيـادة أو نقصـان أو بـالزّيـادة والنّقصـان معا.مثال :طفل اطفال، نمر نم ور، شمس شم وس...

ملاحظة :لقد تم الأخذ بعين الاعتبار القواعد التي تكون متضـمنة أو ملتصـقة بالكلمة وكذلك التي تكون قابلة للبرمجة.

انطلاقـا من النّقـاط المتَوصـل إليهـا من التّحليـل والاقتراحـات السّـابقـة، أصـبحـت الخوارزميّـة تعمـل بـالمنهجيّـة المعتمـدة على الجـذور .وتوضـح الخوارزميّـة المواليّـة مراحـل عمل المجذّع بعد التّحسين:

الخوارزمية: Algorithm Arabic-stemmer Root-Based Stemming.
الهدف: استخراج الجذر من كلمة عربية.
المدخل: كلمة عربية.
المخرج: جذر/احسن جذع ممكن.
الإجراء:
1. الترميز باستعمال Unicode وترجمة الحروف Romanization.
2. التخمين على نوع الكلمة (اسم/فعل) وحالة التعريف.
3. تطبيع أولي للكلمة.
4. حذف اللواحق إن وجدت، وذلك حسب نوع الكلمة.
5. حذف السوابق إن وجدت، وذلك حسب نوع الكلمة.
6. تطبيع نهائي للكلمة.
7. إرجاع أحسن جذع ممكن.
8. مطابقة مع الأوزان ذات طول 6 و5 و4 واستخراج جذر بطول 3.
9. تطبيع نهائي للكلمة.
10. إرجاع النتيجة.

خوارزميّة 2 .الطّبعة القائمة على جذور لـArabicStemmer

حيث أن :

– **المـرحـلـة 2 مـن الـخـوارزمـيّـة الجـديـدة** :إضـافـة بعض النّقـاط **قبل** عمليّة التّجذيع وتحديدا :

- طول الكلمة 4 < وتبدأ ب 'بال 'و 'كال ' تعالج كأسماء معرفة؛
- طول الكلمة 3 < وتبدأ ب' لل 'و' ال ' تعالج كأسماء معرفة؛
- طول الكلمة 2 < وتنتهي بتاء مربوطة تعالج كأسماء.
- **المرحلة 8 من الخوارزميّة الجديدة:** جعل المجذّع يتعامل مع الكلمات المشتقّة، والتي تحتوي على الحروف المحشوة، مثلا :كالكلمات النّاتجة عن جموع التّكسير، ومن صياغات أخرى(كاسم الفاعل، واسم المفعول .)...والتي تكون على صيغة الأوزان البسيطة .في هذه الحالة سيتم الانتقال من مستوى الجذع إلى مستوى أعمق منه ألا وهو الجذر.

3.II. تنفيذ الخوارزميّة: من أجل تنفيذ خوارزميّة التّجذيع اعتمدنا على إطار العمل (Snowball framework) [5] (Porter, 2001) الذّي له إمكانيّة ترجمة الخوارزميّة إلى عدّة لغات برمجة :c، java ، jsx ، python. يعد SnowBall بيئة عمل توفر لنا لغة مصغرة طورت من طرف Martin Porter من أجل معالجة سلاسل الحروف وكتابة خوارزميّات التّجذيع .

4.II. مقارنة الخوارزميّات: لكي يتسنى لنا مقارنة مجذّع (ArabicStemmer) بنسخته القديمة وبباقي مجذّعات اللغة العربيّة، سنستعمل مدونتين :

- المدوّنة الذّهبيّة (Benzahia, 2017) (Golden Corpus). طورت من طرف المطور الجزائري بن زاهيّة لخضر لاختبار مجذّعات اللغة العربيّة. تحتوي 1165 كلمة مع جذوعها وجذورها الصّحيحة.
- المدوّنة القرآنيّة (Dukes, 2009). (Quranic Arabic Corpus) تضم 18350 كلمة من العربيّة المستعملة في القرآن الكريم مع جذوعها الصّحيحة وجذورها إن وجدت. تختلف عن المدوّنة الذّهبيّة باستعمالها للعربيّة الكلاسيكيّة مع علامات التّشكيل.

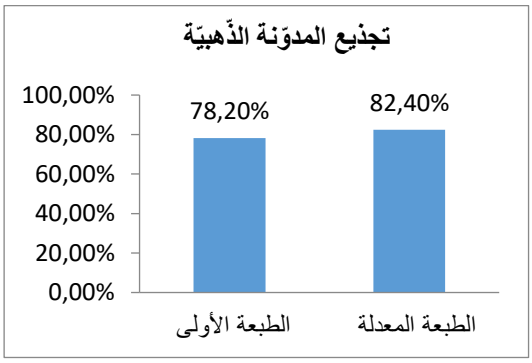
استعملنا سكريبت اختبار بلغة (Python) لإنجاز الاختبارات الضّروريّة المتعلقة بالمجذّع قبل وبعد التّحسين ، وبعد الحصول على نتائج المجذّع لقائمة كلمات المدونتين نقارنها بالنّتائج الصّحيحة وتحسب نسب نجاح الخوارزميّات.

النّتائج Results

بعدما تمت دراسة المجذّع (ArabicStemmer) وفصل مراحله، تمكنا من تسليط الضّوء على مناطقه الرّماديّة حيث عرضنا قائمة نقائصه لتسهيل تعديله وتحسين ادائه. بعدها، تمت إضافة التّحسينات السّالف ذكرها، فأصبح (ArabicStemmer) يعمل على إعطاء أحسن جذع ممكن ويتعامل مع الأوزان البسيطة ويستخرج منها الجذر

.

[5] https://snowballstem.org/



شكل 2. تجذيع المدوّنة الذّهبيّة باستعمال الطّبعتين الأولى والمعدلة لـ ArabicStemmer

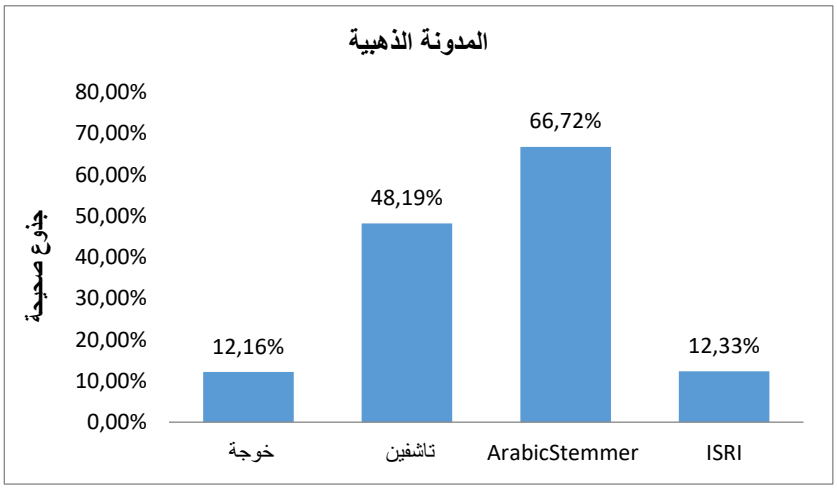
وبالفعل، تمكن المجذّع المُعدَّل من إرجاع الجذوع الصّحيحة للكلمات المُدخلة بزيادة تُقدر ب 4.2 % مقارنة بالطّبعة الأولى كما يبينه شكل 2.

أجرينا بعد هذا، اختبارات حول تجذيع كلمات المدوّنتين الذّهبيّة والقرآنيّة باستخدام مجذّعات خوجة (Khoja, 2003) ، تاشفين (Zerrouki, 2020)، (Chelli, ArabicStemmer 2012) و ISRI (Taghva et al., 2005). والجدير بالذّكر، أنّ خوارزميتيْ ISRI وخوجة تعتمدان التّجذيع القائم على الجذور.

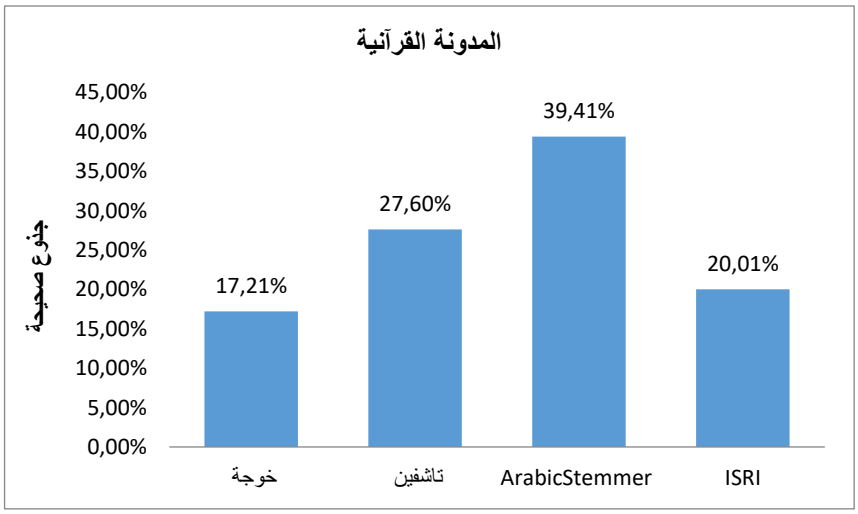
نستطيع اختبار (ISRI) مباشرة على موقع المعالجة المسبقة للنّصوص (text-processing.com/demo/stem)، حيث يوفّر هذا الموقع مجذّعات للغات عديدة. فيما يخص تاشفين، قمنا بإنجاز سكريبت Python وإضافة مكتبة البرنامج 0.3.4.1 (Tashaphyne)(آخر طبعة في وقت كتابة المقال). أمّا مجذّع خوجة فهو متوفّر كمكتبة بلغة Java، متوفّر للتّحميل من الموقع (github.com/motazsaad/khoja-stemmer-command-line) .

تضمّ المدوّنة الذّهبيّة 1165 كلمة أي مايقابل تقريباً 11065 حرفاً. أما عدد كلمات المدوّنة القرآنيّة فيقدر بـ 18350 كلمة ويعادلها 133852 حرفا؛ ويشكل هذا العدد الكبير من الأحرف عائقاً لـــ ISRI وتاشفين. وعليه، يجب تقسيم المدوّنة القرآنيّة إلى 12 جزءًا لا يتعدى عدد أحرف كل جزء 50000 حرفاً.

يبين الشّكلين شكل 3. وشكل 4. نسب استخراج جذوع كلمات المدونتين الذّهبيّة والقرآنيّة باستخدام المجذّعات المذكورة سالفاً.

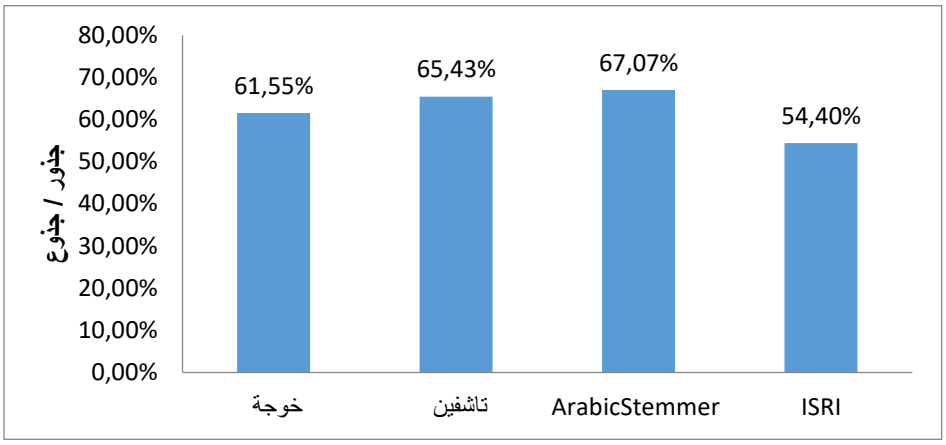


شكل 3. تجذيع المدوّنة الذّهبيّة Golden Corpus

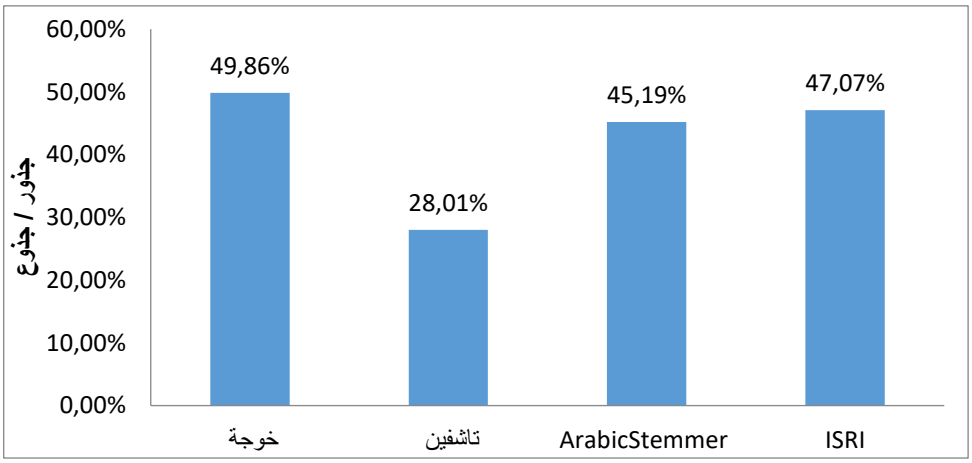


شكل 4. تجذيع المدوّنة القرآنيّة Quranic Corpus

نلاحظ مدى تفوق خوارزميّة (ArabicStemmer)على نظيراتها من مجذّعات اللغة العربيّة حيث أنه قام بتجذيع 774 كلمة من أصل 1165 كلمة في المدوّنة الذّهبيّة و7232 كلمة من أصل 18350 كلمة في المدوّنة القرآنيّة. نلاحظ أيضاً تقارب نسب مجذّع تاشفين وArabicStemmer.



شكل 5. استخراج الجذور / الجذوع من المدوّنة الذّهبيّة



شكل 6. استخراج الجذور / الجذوع من المدوّنة القرآنيّة

والملاحظ في شكل5. والمتعلّق بأداء المجذّعات في استخراج الجذوع والجذور، فإن النّسب تتقارب فيما يخص المدوّنة الذّهبيّة؛ أما فيما يتعلّق بالمدوّنة القرآنيّة أن مجذّع خوجة يتفوّق بنسبة 49.86% على الرّغم من عدم تجذيعها لـ 304 كلمة، يليها ISRI بنسبة 47.07% وبعدها (ArabicStemmer) بنسبة 45.19%؛ كما هو مبيّن في شكل6.

وإذا نظرنا عن قرب لبعض الجذوع المستخرجة، فنجد المخرجات المبينة في **جدول 1.**

جدول 1. بعض الكلمات من المدوّنة الذّهبيّة والقرآنيّة مع جذوعها وجذورها الصّحيحة والمستخرجة

الكلمة	الجذر	الجذع	خوجة مجذّع	مجذّع تاشفين	ArabicStemmer	مجذّع ISRI
استنتجتاه	نتج	استنتج	استنتجتاه	**استنتج**	**استنتج**	استنتجتاه
فأطفالهم	طفل	اطفال	<u>طفل</u>	طفال	**اطفال**	أطفال
مدرسات	درس	مدرس	<u>درس</u>	**مدرس**	**مدرس**	<u>درس</u>
تعضلوهن	عضل	تعضل	تعضلوهن	تعضلوهن	**تعضل**	عضلو
لأماناتهم	امن	أمانات	لأم	لاماناتهم	لامان	لأم

ملاحظة: الكلمات بالخطّ العريض هي جذوع صحيحة، والتي تحتها سطر هي جذور صحيحة، أمّا الباقي فيعتبر خطأ في التّجذيع.

يتبيّن من الجدول، أن(ArabicStemmer) نجح في التّعرّف على الكلمات وفصلها عن حروف العطف والضّمائر عدا "لأماناتهم"، ونجح في تجذيع جمع المؤنث السّالم "مدرسات" لكنه أَخْفَقَ في "لأماناتهم".

المناقشة Discussion

من حصولنا على نتائج تنفيذ خوارزميّات تجذيع اللغة العربيّة، وعلى رأسها المجذّع (ArabicStemmer) بنسختيه، يمكننا مناقشة ذلك بالتّرتيب فيما يلي :

أولاً: أن تفوق النّسخة المعدلة على النّسخة الأولى بنسبة 4.2% (أنظر شكل 2.) في تجذيع كلمات المدوّنة الذّهبيّة، يمكن ربطه بتحسين المرحلة 2 والمتعلقة بالتّخمين على نوع الكلمة (إسم أو فعل) . حيث أن المجذّع أصبح يعالج الأسماء التي يزيد طولها عن 4 وتنتهي بـ "ات" أو "تين/تان" أو في اخرها خفض أو تنوين وهذا ما زاد من فعاليته. وعليه، قادتنا هاته النّسب إلى اعتماد النّسخة المُعدَّلة في المقارنة مع باقي المجذّعات.

ثانياً: أن نسبة استخراج الجذوع من المدوّنة الذّهبيّة فاقت النّصف بالنّسبة لـ (ArabicStemmer) حيث جذَّع 66.72% من كلمات المدوّنة أي ما يعادل 774 كلمة من 1165 كلمة. في حين أن مجذّع تاشفين (Zerrouki, 2020) جَذَّع ما يقارب النّصف أي 48.19%. لكن نجد أن مجذّع خوجة (Khoja, 2003) و ISRI (Taghva et al., 2005)استخرجا ما يعادل الثُّمُن $\left(\frac{1}{8}\right)$ فقط من الكلمات الموجودة في المدوّنة الذّهبيّة (أنظر شكل3.). وإذا نظرنا عن كثب إلى هذين المجذّعين، نجد أنهما يعتمدان على استخراج الجذور بدل الجذوع وهذا ما يفسر تدنّي نِسَبِهم.

أمَّا فيما يتعلق بالمجذّع (ArabicStemmer)، فتَفَوُقُه راجع إلى عدّة نقاط :

- نظام التّهجئة المتبع والذّي يقودنا إلى استخلاص جذوع صحيحة، عكس الأنظمة الأخرى، مثال: مجذّع خوجة يستبدل حروف العلة بحرف 'الواو' مما يؤدي إلى تجذيع مغاير تماماً للجذع الصّحيح.
- التّطبيع النّهائي والذّي يُعنى بتطبيع الهمزات على جميع أشكالها : بئس=> بيس ، إسلام => اسلام، تأثر=> تاثر، ...
- معالجة أ الاستفهام، مثال: «أأباد؟ " حيث جذعها (ArabicStemmer) " اباد" وهو الجذع الصّحيح، أما خوجة وISRI فجذعاها "أبد " في حين أن تاشفين استخرج كلمة "باد" كجذع. أما الجذر الصّحيح فهو "بيد".
- تطبيع الهاء والواو، مثال: "استفتيتموه" ، حيث أن (ArabicStemmer) وتاشفين استخرجا الجذع الصّحيح "استفتي" أما خوجة وISRI عجزا عن تجذيعه .

ثالثاً: أن، كما يتبين من شكل4. ، نسبة استخراج الجذوع من المدوّنة القرآنيّة Quranic Corpusباستعمال (ArabicStemmer) تعادل 39.41% من كلمات المدوّنة ويتفوق بذلك على باقي المجذّعات رغم أن العربيّة المستعملة هي عربيّة تقليديّة ومُشَكَّلة. ويعود ذلك إلى:

- فعاليّة التّخمين على نوع الكلمة إسم/فعل، إذ نلاحظ أن (ArabicStemmer) نجح في التّعرف على نوع الكلمات، مثال: "فذبحوها" و "يعتدون" و "هادوا" على أساس أنها **أفعال** وجذَّعها " ذبح " و "يعتد" و "هاد" في

– حين عجزت باقي المجذّعات عن استخراج الجذوع /الجذور الصّحيحة .

– التّعرف على الأوزان البسيطة، مثال: "ووالدّ" على أساس "وفاعل"

رابعاً : أنه إذا تم الأخذ بعين الاعتبار الجذور والجذوع في آنٍ واحدٍ، فإن آداء الخوارزميّات الأربع يفوق النّصف بنسب متقاربة كما يوضحه الشّكلان شكل5. وشكل 6. وهذا ما يُرجع مجذّعيْ خوجة وISRI إلى منصة التّتويج، حيث أنهما يعتمدان على استخراج الجذور بدل الجذوع. ولكل نوع استخداماته، فالجذر مثلا يغير معنى الكلمة المستخرجة، كإسلام وسُلَّم لهما نفس الجذر "سلم".

خامساً: أن المقارنة المطلقة للمجذّعات قد تكون غير دقيقة، وذلك يعود إلى اتباع كل خوارزميّة لمنطق مختلف في مراحل معالجة النّصوص ككيفيّة الفصل إلى كلمات، واستبعاد كلمات الوقف، وطريقة التّطبيع والتّرميز وما إلى ذلك. كل هذا بسبب غياب معايير موحدة في التّطبيع العربي، مما يؤثر على نتيجة المجذّع . وهذا ما أدى بمجذّع خوجة إلى إنتاج جذورًا ليس لها علاقة بالكلمة الأصليّة، فمثلا كلمة منظمات جُذِّعت إلى "ظما"، في حين كان يجب أن تُجذَّع إلى "نَظم" وذلك لأنه يقوم باستبدال حروف العلة التي تظهر في الجذر بحرف 'الواو'.

سادساً: أن معظم المجذّعات التي تستخدم قواميس تتطلب تحديثا مستمرا للصيانة .قد يكون هذا، عاملاً في نجاح الخوارزميّة في استنباط الجذور أو الجذوع الصّحيحة إذا ما كانت محفوظة في القواميس. بدليل، مجذّع خوجة وتفوّقه على البقيّة في المدوّنة القرآنيّة. لكنه وفي نفس الوقت، يشكل عائقاً في تجذيع الكلمات غير المُخزَّنة أو المكتشفة حديثاً.

سابعاً: وتأتي هذه النّتائج لتؤكد ما توصّل إليه القحطاني وفريقها في (Alqahtani *et al.,* 2020) ، حيث أن دراستهم خلُصت إلى أن (ArabicStemmer) دقيق بنسبة 86% في تجذيع 6000 كتاب من المكتبة الشّاملة[6] مقارنة بـ 81% لمجذّع خوجة. وبالمقابل، تتنافى نتائجنا مع نتائج المزروع وفريقها (Almazrua *et al.,* 2020) حيث توصلوا إلى أن مجذّعي خوجة وISRI احسن من (ArabicStemmer) في تجذيع مجموعة البيانات (التّغريدات) EveTar[7] (Hasanain *et al.,* 2018).

ثامناً: أنه وبالرّغم من التّحسينات المنجزة إلا أنه تبقى هناك الكثير من التّعديلات التي يمكن إضافتها، ومن أهمها:

– الأوزان التي تحتوي على لام ثانيّة.

– الجذور المعتلة.

– معالجة أوزان جموع التّكسير التي لها أكثر من صيغة في المفرد.

الخلاصة Conclusion

من خلال عملنا هذا، وكإجابة على الإشكاليّة المطروحة في بدايّة البحث، وبعد مناقشة النّتائج المحصلة، توصلنا إلى خلاصة يمكن اختصارها فيما يلي:

– تحسين أداء الخوارزميّة بنسبة 4.2% وذلك بتحسين مرحلة التّخمين على نوع الكلمة باعتبار الخفض والتّنوين من علامات الاسم

[6] https://shamela.ws/

[7] http://qufaculty.qu.edu.qa/telsayed/evetar/

– وسين الاستقبال وتاء التّأنيث من علامات الفعل.

– قيام مجذّع (ArabicStemmer) باستخراج 39.41% من الجذوع الصّحيحة من أصل 18350 كلمة في المدوّنة القرآنيّة، على الرّغم من كونها مكتوبة بالشّكل التّقليدي للحروف مع علامات التّشكيل. ويعود هذا إلى مرحلتَيْ التّطبيع الأوّلي والنّهائي وتعاملهما مع الهمزات بشكل أدى الى التّجذيع الصّحيح.

– وبالنّسبة للّغة المعاصرة، فقد تعامل معها المجذّع بشكل جيّد حيث تفَوَّق على ثاني أحسن مجذّع بنسبة 18.53% في المدوّنة الذّهبيّة بتجذيعه لـ 66.72% من الكلمات. كل هذا بفضل عدم استعمال قواميس معدّة مسبقا والتي تشكل عقبة في وجه المجذّعات، وكذلك بفضل مرحلة التّعرف على الأوزان البسيطة وجموع التّكسير.

– إضافة إلى هذا، قمنا بإدراج المجذّع ضمن مكتبات معالجة اللغة الطّبيعيّة.

وعليه نكون قد أسهمنا بعض الشّيء في رفع التّحديّات المذكورة في هذا العمل، كالتّعامل مع جمع المؤنث السّلم وجمع التّكسير والأسماء المنونة. ورغم ذلك، تبقى التّحديّات مرفوعة في هذا المجال كمعالجة الأفعال المعتلة والأوزان التي تضم لامين وجمع التّكسير بأكثر من صيغة مفرد وغيرها من خصائص اللغة العربيّة التي تنفرد بعذوبة وغنًى لامتناهيين.

المراجع References

Alqahtani, A., Alnefaie, M., Alamri, N., & Khorsi, A. (2020): Enhancing the Capabilities of Solr Information Retrieval System: Arabic Language. *2020 3rd International Conference on Computer Applications Information Security (ICCAIS)*, 1–5. https://doi.org/10.1109/ICCAIS48893.2020.9096810

Benzahia, L. (2017): *Golden Arabic Corpus* [Python]. Ibn Malik Arabic NLP tools. https://github.com/ibnmalik/golden-corpus-arabic

Chelli, A. (2012): *Proposal of an Advanced Retrieval System for Noble Qur'an* [Magister]. Ecole Nationale Superieure en Informatique.

Dawson, J. L. (1974): Suffix removal and word conflation. *Bulletin of the Association for Literary and Linguistic Computing*, *2*(3), 33–46.

Dukes, K. (2009): The Quranic Arabic Corpus. *Language Research Group University of Leeds*. https://corpus.quran.com/

Hasanain, M., Suwaileh, R., Elsayed, T., Kutlu, M., & Almerekhi, H. (2018): EveTAR: Building a large-scale multi-task test collection over Arabic tweets. *Information Retrieval Journal*, *21*(4).

Taghva, K., Elkhoury, R., & Coombs, J. (2005): Arabic stemming without a root dictionary. *International Conference on Information Technology: Coding and Computing (ITCC'05)-Volume II*, *1*, 152–157.

Zerrouki, T. (2020): *Towards An Open Platform For Arabic Language Processing* [Ph.D.]. Ecole Nationale Superieure en Informatique.

الأنباري, ك. ا. (1995): *أسرار العربيّة*. دار الجيل.

بن آجروم, م. (2002): *متن الآجُرّوميّة*. مكتبة الآداب للطباعة والنّشر والتّوزيع أول طبعة 1814.

بن جنّي, أ. ا. (1988): *اللمع في العربيّة*. دار مجدلاوي للنّشر والتّوزيع.

روينة, إ. (2019): كلمة رئيس التّحرير. *مجلة العلوم والتّكنولوجيا, المجلس الأعلى للغة العربيّة, بالجزائر, 1*. https://www.asjp.cerist.dz/en/PresentationRevue/714

زلاقي, ع. (2018): التّرجمة الآليّة للغة العربيّة—التّحديّات والآفاق لحوسبة اللغة العربيّة-. *اللغة العربيّة والتّقانات الجديدة, 1*, 423–455.

https://doi.org/10.1007/s10791-017-9325-7

Khoja, S. (2003): *APT: An automatic Arabic part-of-speech tagger* [Ph.D., Lancaster University]. https://eprints.lancs.ac.uk/id/eprint/12350/

Larkey, L. S., Ballesteros, L., & Connell, M. E. (2007): Light Stemming for Arabic Information Retrieval. In A. Soudi, A. van den Bosch, & G. Neumann (Eds.), *Arabic Computational Morphology: Knowledge-based and Empirical Methods* (pp. 221–243). Springer Netherlands. https://doi.org/10.1007/978-1-4020-6046-5_12

Lovins, J. B. (1968): Development of a stemming algorithm. *Mech. Transl. Comput. Linguistics*, *11*(1–2), 22–31.

Manning, C. D., Raghavan, P., & Schütze, H. (2010): Introduction to information retrieval. *Information Retrieval*, *13*(2), 192–195. https://doi.org/10.1007/s10791-009-9115-y

Porter, M. F. (1980): An algorithm for suffix stripping. *Program*.

Porter, M. F. (2001). *Snowball: A language for stemming algorithms*.