خوارزمية مقترحة لتحسين استخراج المقاطع النصيّة في نظم إجابة الأسئلة بالعربي

م.لانا الصّباغ[1]، د.أميمة الدكاك[2]، د.ندى غنيم[3]

الملخص

تعتبر نظم الإجابة الآلية اليوم بالغة الأهمية، ويعود ذلك لحجم المعطيات الكبير المتاح على الإنترنت، وبالتالي أصبح المستخدمين بحاجة نظم قادرة على تقديم إجابات مختصرة ودقيقة على استفساراتهم من ضمن هذه المعطيات، وتعد مرحلة استخراج المقاطع النصيّة من أهم المراحل في نظم الإجابة الآلية، وذلك لأن الحصول على إجابات صحيحة مرتبط باستخراج مقاطع نصيّة صحيحة، لذلك، ارتكز بحثنا على إيجاد خوارزمية لتحقيق مرحلة استخراج المقاطع النصيّة تُحسن من أداء نظم الإجابة الآلية بشكل عام. قمنا باقتراح خوارزمية تعتمد على دمج التشابه النحوي والدلالي ما بين نص السؤال والمقاطع النصيّة، حيث جرى قياس التشابه النحوي باستخدام خوارزمية BM25، وجرى قياس التشابه الدلالي باستخدام تقنية تضمين الكلمات وباعتماد نموذج AraVec المدرب مسبقاً. جرى اختبار الخوارزمية المقترحة باستخدام مجموعة بيانات منمّطة ARCD والمستخدمة ذاتها في أبحاث مشابهة سابقة، واستطعنا الحصول على مقاطع نصيّة صحيحة بدقة 92.4%، بحيث يعتبر المقطع النصي صحيح في حال احتوى على الإجابة الصحيحة المرفقة في عينة الاختبار.

الكلمات المفتاحية: معالجة اللغات الطبيعية، الإجابة الآلية، استخراج المقاطع النصيّة.

(1) طالب ماجستير، نظم معطيات كبيرة، المعهد العالي للعلوم التطبيقية والتكنولوجيا.
(2) رئيس قسم الاتصالات للشؤون التعليمية، المعهد العالي للعلوم التطبيقية والتكنولوجيا.
(3) استاذ مساعد، كلية الهندسة المعلوماتية، جامعة دمشق.

A **Suggested Algorithm to Improve Passage Retrieval in Arabic Question Answering Systems**

Eng. Lana AlSabbagh[4], Dr. Oumayma AlDakkak[5], Dr. Nada Ghneim[6]

Abstract

Users often have specific questions in mind, for which they hope to get answers. Question Answering Systems (QAS) aim at retrieving accurate answers for the user's questions from a large dataset. Passage retrieval is a crucial component for any QAS, it identifies top-ranked passages that may contain the answer for a given question. Also, it is a longstanding challenge widely studied over the last decades. However, it still requires further efforts in Arabic QAS. In this research, we focus on the passage retrieval phase to get the most related passages to the correct answer. We suggested a model that measures the similarity between passages and the question and combines the BM25 ranker and Word Embedding approach. We tested our system on the ACRD dataset, the system was able to achieve an accuracy of 92.4% in finding the passages that contain the correct answer for a given question.

Keywords: Natural Language Processing, Question Answering, Passage Retrieval.

[4] Master Student, Big Data, Higher Institute for Applied Science and Technology (HIAST).
[5] Head of Communications Department for Educational Affairs, Higher Institute for Applied Science and Technology (HIAST).
[6] Assistant Professor, Faculty of Information Technology Engineering, Damascus University.

1- المقدمة

الإجابة الآلية هي فرع من فروع معالجة اللغات الطبيعية التي تهتم باستخراج إجابات دقيقة ومختصرة عن أسئلة المستخدمين، بدلاً من الاقتصار فقط على أكثر المستندات المرتبطة بهذه الأسئلة [1]. وطُرح في الفترة الأخيرة العديد من الدراسات التي عُنيت بهذا المجال ومنها ما حقق نتائج جيدة جداً. ولمّا كان هذا التطور منتشراً بكثرة للغات الأجنبية ومحدوداً للغة العربية، ومع انتشار الثقافة التقنية العربية، كان لابد من ظهور تقنيات داعمة للغة العربية، ومن هنا تمهّد الطريق أمام بحثنا لإيجاد خوارزميات تُحسّن من أداء نظم الإجابة الآلية باللغة العربية.

تختلف مكونات نظم الإجابة الآلية باختلاف الغرض منها، هناك بعض النظم المتخصصة في الإجابة عن أسئلة ضمن مجال محدد، مثال الإجابة عن الأسئلة الإسلامية فقط [2]، وهناك نظم إجابة آلية متخصصة فقط بالإجابة عن أسئلة من أنماط محددة، مثال النظم التي تجيب عن أسئلة من نوع "لماذا" فقط [3]. توجد نظم إجابة آلية تهتم بالإجابة عن الأسئلة المطروحة ضمن مجالات متعددة (غير مقيدة بمجال محدد)، تدعى مثل هذه النظم "نظم الإجابة الآلية مفتوحة المجال"، وهي النظم الأكثر شمولية في مجال الإجابة الآلية [4,5].

تمر نظم الإجابة الآلية عادةً بأربعة مراحل رئيسية وهي على التالي: مرحلة معالجة السؤال، ومرحلة استخراج المستندات، ومرحلة استخراج المقاطع النصيّة، ومرحلة استخراج الإجابة. بحيث تُطبق في مرحلة معالجة السؤال خطوات المعالجة اللغوية على نص السؤال، لتجري مطابقته لاحقاً مع المستندات النصيّة التي تشكل مجموعة بيانات النظام، وذلك ضمن مرحلة استخراج المستندات، والتي تعيد أكثر المستندات ارتباطاً بنص السؤال (يمكن تحديد أفضل n مستنداً)، يجري بعد ذلك استخراج المقاطع النصيّة الأكثر ارتباطاً بالسؤال وذلك ضمن مرحلة استخراج المقاطع النصيّة، وأخيراً استخراج الإجابة الدقيقة والمحددة للسؤال من ضمن المقاطع النصيّة الناتجة، وذلك ضمن مرحلة استخراج الإجابة [6]. تعد مرحلة استخراج المقاطع النصيّة من المراحل الهامّة في نظم الإجابة الآلية، ويعود ذلك لكون الحصول على إجابة صحيحة مرتبط بظهور هذه الإجابة ضمن المقاطع النصيّة الناتجة عن هذه المرحلة، وبالتالي تنعكس كفاءة نظم الإجابة الآلية وبشكل كبير على كفاءة هذه المرحلة. تُنفذ مرحلة استخراج المقاطع النصيّة عادةً باستخدام طرق قياس التشابه التقليدية، مثل نموذج فضاء المتجهات Vector Space Model وهو نموذج جبري يستخدم لتمثيل مستندات نصيّة كمتجهات رقمية، وتشكل مجموعة المصطلحات أبعاد هذه المتجهات بحيث يمثل كل مصطلح بعد، ويأخذ المتجه قيمه وفقاً لعدة آليات تعتمد على معلومة تكرار المصطلح ضمن المستندات، ومن أحد الطرق المعتمدة لحساب قيم هذه المتجهات Term Frequency Inverse Document Frequency(TF-IDF)، حيث يجري حساب هذا المقياس بحاصل جداء عدد المرات التي يظهر فيها المصطلح t في المستند d وتردد المستند العكسي للمصطلح t عبر مجموعة المستندات D كما توضح المعادلة التالية [13]:

$$tf\ idf\ (t, d, D) = tf\ (t, d)\ .\ idf\ (t, D) \quad (1)$$

ويجري حساب tf و idf وفقاً للمعادلات التالية:

$$tf\ (t, d) = \log\ (1 + freq\ (t, d)) \quad (2)$$

$$idf\ (t, D) = \log\ \left(\frac{N}{count\ (d \in D : t \in d)} \right) \quad (3)$$

بحيث:

- freq(t,d) تمثل عدد مرات تكرار المصطلح t ضمن المستند d.
- N عدد المستندات في مجموعة المستندات D.

وهناك طرق أخرى لقياس التشابه تُستخدم عادةً في مرحلة استخراج المقاطع النصيّة مثل خوارزمية Best Matching 25(BM25)، وهي خوارزمية تستخدمها محركات البحث لترتيب مجموعة من المستندات وفقاً لمدى صلتها باستعلام بحث معين، حيث تقاس مدى صلة مستند ما D باستعلام معين Q وفقاً للمعادلة التالية [14]:

$$\text{score}(D,Q)=\sum_{i=1}^{n}\text{IDF}(q_i)\cdot\frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot\left(1-b+b\cdot\frac{|D|}{\text{avgdl}}\right)} \quad (4)$$

بحيث:

- $f(qi,D)$ قيمة تكرار المصطلح qi ضمن المستند D.
- $|D|$ طول المستند D مقاساً بعدد الكلمات.
- $avgd1$ متوسط طول المستند في المجموعة النصيّة التي يتم سحب المستندات منها.
- $k1,b$ معاملات حرة تأخذ عادةً القيم التالية وفقاً للتجريب والاختبار:

$$k_1 \in [1.2, 2.0]$$

$$b = 0.75$$

- $IDF(qi)$ وهو قياس تردد المستند العكسي من أجل مصطلح معين qi والذي يتم حسابه وفقاً للمعادلة رقم (3).

تعتمد مثل هذه الطرق (نموذج فضاء المتجهات وBM25) على قياس مدى التشابه بين السؤال والمقاطع النصيّة من الناحية النحوية فقط، أي لا تأخذ الدلالة بالاعتبار أثناء قياس التشابه، لهذا السبب يمكن تعزيز القياس السابق بقياس آخر يأخذ الدلالة بالاعتبار، واعتمدت الدراسات المطروحة في مجال قياس التشابه بين النصوص على قياس التشابه الدلالي باستخدام الشبكات الدلالية[7] Wordnets، ولكن تعاني أساليب قياس التشابه باستخدام الشبكات الدلالية أنها لا تنظر لسياق الكلمة عند حساب التشابه، كما في مثال كلمتي "فأرة" و"قطة"، تعتبر هاتان الكلمتان متشابهتان من الناحية الدلالية باعتماد الشبكات الدلالية لكن في حال وردت كلمة "فأرة" ضمن السياق التالي "اشتريت اليوم فأرة لحاسبي المكتبي"، نلاحظ أن المعنى هنا أصبح مختلف تماماً، وبالتالي يمكن للكلمة أن تختلف دلالتها باختلاف السياق الواردة ضمنه، لذلك، لابد من اعتماد أساليب جديدة في قياس التشابه الدلالي تأخذ سياق الكلمة بالاعتبار أثناء حساب قيمة التشابه.

ظهرت مؤخراً تقنيات حديثة في مجال معالجة اللغات الطبيعية، مثل تقنية تضمين الكلمات Word Embedding وهي تقنية تعتمد على تمثيل الكلمة بمتجه رقمي، ويجري إنشاء هذه المتجهات اعتماداً على مدونات نصيّة كبيرة، وباستخدام شبكات عصبونية تأخذ السياقات المختلفة للكلمة من المدونات النصيّة وتعطي المتجه الرقمي الممثل للكلمة [9,10]. تمتلك الكلمات المتشابهة في سياقها تمثيلاً متقارباً أيضاً، كما في الجملتين "يعمل أبي أستاذ في المدرسة" و"يعمل أبي مُدرس في المدرسة". تستخدم تقنية تضمين الكلمات في العديد من مهام معالجة اللغات الطبيعية مثل تحليل المشاعر ومحركات التوصية Recommendation Engines ونظم استرجاع المعلومات وغيرها الكثير، وأثبتت هذه التقنية كفائتها في مثل هذه التطبيقات؛ وبالتالي يمكن الاستفادة منها للتحسّين من أداء عملية استرجاع المقاطع النصيّة في نظم الإجابة

[7] **الشبكة الدلالية** هي شبكة تمثل علاقات دلالية بين المفاهيم، وغالباً ما تستخدم كطريقة لتمثيل المعرفة .تعتبر الشبكة مخطط موجهاً أو غير موجه مؤلفاً من عقد، التي تمثل المفاهيم بالإضافة إلى الخطوط.

الآلية، وذلك باستخدامها في حساب التشابه الدلالي بين نص السؤال والمقطع النصي المُستخرج، وهذا ما قمنا بتجريبه واختباره في بحثنا.

نقترح في هذا البحث خوارزمية لتنفيذ مرحلة استخراج المقاطع النصيّة في نظم الإجابة الآلية باللغة العربية، بحيث قمنا نتيجة التجريب والاختبار بوضع نموذج خطي لقياس التشابه بين السؤال والمقطع النصي. يجمع النموذج المقترح بين قياسين التشابه النحوي والدلالي، وذلك اعتماداً على تقنية تضمين الكلمات لقياس التشابه الدلالي، وخوارزمية BM25 لقياس التشابه النحوي.

سنتابع في تتمة هذه الورقة البحثية بدراسة لأهم الأبحاث والدراسات المشابهة الحديثة ضمن نفس المجال في قسم الدراسة المرجعية. نشرح في قسم مجموعة البيانات المعطيات المستخدمة في وضع الخوارزمية المقترحة واختبارها، ونعرض في قسم الخوارزمية المقترحة الطريقة المقترحة مع شرح للمراحل التي تمر بها، قمنا في قسم النتائج باختبار المنهجية المقترحة وعرض النتائج ومقارنتها بالأعمال المشابهة، وعرضنا في قسم الأدوات المستخدمة المكاتب والأدوات التي قمنا بالاستعانة بها لتحقيق الخوارزمية المقترحة، ونختم في قسم الخلاصة بعرض ملخصاً عن البحث والآفاق المستقبلية له.

2- الدراسة المرجعية

هناك العديد من الأبحاث التي نُشرت في مجال الإجابة الآلية باللغة العربية، ومنها ما اهتم بالإجابة عن أسئلة ضمن مجال محدد، مثال نظام ASHLK، وهو نظام متخصص في الإجابة عن الأسئلة المطروحة ضمن الأحاديث النبوية الشريفة [2]، يسترجع النظام جملاً مقتطعة من الأحاديث النبوية الشريفة، والأكثر ارتباطاً بالسؤال المطروح كإجابة عن هذا السؤال، ويعتمد في إيجاد الإجابة على قياس التشابه بين السؤال المطروح وجمل الأحاديث النبوية، وذلك باستخدام نموذج التشابه باعتماد البيان Graph Similarity، بحيث تمثل كل عقدة في البيان جملة نصية وتمثل كل رابط في البيان قيمة التشابه بين الجملتين (العقد)، يقاس التشابه من خلال الدمج بين التشابه النحوي بين كلمات الجملتين بالإضافة إلى التشابه الدلالي في قياس واحد، حيث جرى قياس التشابه النحوي اعتماداً على مدى تطابق جذوع كلمات الجملتين، وجرى قياس التشابه الدلالي باستخدام الشبكات الدلالية، بحيث تُعتبر جملتين متشابهتين دلالياً وفقاً لعدد المرادفات المشتركة بينهما والمستخرجة من الشبكات الدلالية. يُأخذ ترتيب الكلمات في الجملة بالاعتبار أثناء قياس التشابه النحوي (على مبدأ أن ترتيب الكلمات في الجملة يعكس معنى هذه الجملة)، يعيد النظام عدداً محدداً من الجمل التي تكون أكثر تشابهاً مع بعضها البعض وأكثر تشابهاً مع سؤال المستخدم كإجابة، وذلك بعد حذف الجمل المتكررة (التشابه فيما بينها كبير جداً). جرى اختبار النظام على مجموعة بيانات تتكون من 4000 حديث و3825 سؤال، مقسمة إلى 2678 سؤال من أجل مرحلة التدريب، و1147 سؤال من أجل الاختبار، وكانت دقة الحصول على إجابة صحيحة 83.5٪.

هناك نظم إجابة آلية أخرى اهتمت بالإجابة عن أسئلة من نمط محدد، مثال نظام Lemaza، وهو نظام إجابة آلية باللغة العربية مفتوح المجال، ومتخصص بالإجابة عن أسئلة من نمط "لماذا" [3]، يمر النظام بالمراحل الرئيسية الأربع. تجري المراحل الثلاثة الأولى وفقاً للآليات المستخدمة عادةً في نظم الإجابة الآلية، حيث جرى تطبيق خطوات المعالجة اللغوية على نص السؤال في مرحلة تحليل السؤال، وتطبيق تقنيات استرجاع البيانات التقليدية في تحقيق مرحلتي استخراج المستندات واستخراج المقاطع النصيّة، وينحصر الاهتمام في هذا البحث بمرحلة

استخراج الإجابة، حيث يجري ضمن هذه المرحلة استخراج العبارات الجديلة Cue Phrases، وهي العبارات التي تحوي جملاً مرتبطة بأدوات ربط من نوع (بسبب، نتيجة لذلك، لام السببية، علاوة على ذلك...)، ويجري تحديد الجملة المناسبة وفقاً لنوع السؤال المطروح (تعليل، تفسير...) وإعادتها كإجابة عن السؤال. جرى اختبار النظام على مجموعة بيانات تم جمعها بشكل يدوي من قبل الباحثين، ومؤلفة من 110 سؤال، استطاع النظام أن يجيب بشكل صحيح بدقة 79.2%.

يوجد نظم إجابة آلية باللغة العربية اهتمت بالإجابة عن الأسئلة المطروحة بمجالات متعددة، وغير مقيدة بأنماط أسئلة محددة، مثال نظام SOQAL، وهو نظام إجابة آلية باللغة العربية معتمد على النماذج اللغوية العصبونية [11]. تمر عملية استخراج الإجابة أولاً بمعالجة نص السؤال وفقاً لخطوات المعالجة اللغوية التقليدية وهي: تقطيع السؤال إلى كلمات، وإزالة الأحرف والكلمات غير المنتمية لأبجدية اللغة العربية، واستنظام نص السؤال[8]، وتجذيع كلمات السؤال (باستبدال كل كلمة بالجذع). يجري لاحقاً استخراج المستندات الأكثر ارتباطاً بالسؤال المطروح من ضمن مجموعة البيانات المعتمدة باستخدام Vector Space Model مع قياس TF-IDF،كما توضح المعادلات (1) و(2) و(3)، ومن ثم تجري عملية استخراج المقاطع النصيّة بقياس التشابه بين الكلمات الواردة في المستند مع كلمات السؤال، ويقاس التشابه باعتماد نموذج عصبوني مدرب مسبقاً BERT (Bidirectional Encoder from Representations Transformers)، بحيث تعتبر الكلمات متشابهة بمدى ورودها في نفس السياق عادةً، وبذلك يجري اقتطاع المقطع النصي الذي يحوي أكثر الكلمات تشابهاً مع نص السؤال وبطول 10 كلمات كحد أعظمي، ويقدمه النظام كإجابة نهائية عن السؤال المطروح. جرى تدريب النظام على مجموعة معطيات Arabic-SQuAD واختباره على مجموعة بيانات تم إعدادها من قبل الباحثين ونشرها باسم Arabic Reading Comprehension Dataset (ARCD)، وهي مجموعة بيانات منمّطة ومؤلفة من 1935 سؤالاً. استطاع النظام أن يجيب بشكل صحيح عن الأسئلة المطروحة بدقة 91.4%، بحيث يعتبر الجواب صحيحاً في حال كانت الإجابة الصحيحة المرفقة في مجموعة البيانات محتواة ضمن المقطع النصي الذي يعيده النظام.

3- مجموعة البيانات

اعتمدنا في بحثنا على مجموعة البيانات Arabic Reading Comprehension Dataset[9] (ARCD)، والمعتمدة ذاتها ضمن البحث SOQAL [11]، وقمنا باختيارها كمجموعة معطيات للنظام نظراً لشموليتها من حيث تنوّع المجالات التي تتناولها، وتنوّع أنماط الأسئلة المطروحة ضمنها. تتألف مجموعة البيانات هذه من 1395 ثنائية (سؤال، وجواب)، بحيث يتمثل الجواب بجزء نصي مقتطع من المستند الذي طُرح منه السؤال، جرى استخراج هذه الثنائيات (الأسئلة والأجوبة) من 465 مستند نصي ضمن 155 مجالاً مختلفاً، وذلك بمعدل استخراج ثلاثة أسئلة من المستند النصي الواحد على شكل ثنائية (مقطع نصي، وسؤال ضمن المقطع النصي). جرى في بحثنا تقسيم مجموعة البيانات هذه بنسبة 40% لتدريب

[8] **استنظام النص Text Normalization:** وهي عملية تحويل النص لصيغة موحدة بحيث يمكن معالجته لاحقاً، ويتم ذلك بتوحيد كتابة بعض الأحرف (مثال توحيد أ، إ، ا، آ لتكتب فقط كـ "ا").

[9] https://github.com/husseinmozannar/SOQAL/tree/master/data

النظام المقترح واعتماد الخوارزميات المستخدمة ضمنه و60% لاختبار النظام.

4- الخوارزمية المقترحة

لاحظنا من خلال الدراسة المرجعية التي قمنا بها بأن أغلب نظم الإجابة الآلية التي اهتمت باللغة العربية كانت محدودة إما بمجال المعطيات (محددة المجال)، مثال نظام ASHLK، أو بأنماط الأسئلة التي تعالجها، مثال نظام Lemaza، كما أن المعالجة المتّبعة في قياس التشابه الدلالي ضمن هذه الأبحاث ارتكزت على أساليب قديمة نوعاً ما ولاتأخذ سياق الكلمات بالاعتبار. ولمّا ظهرت أبحاث تهتم بأنماط مختلفة من الأسئلة ومفتوحة المجال وتستخدم تقنيات حديثة في حساب التشابه الدلالي؛ اقتصرت على إعادة جزء نصي قصير كإجابة نهائية عن السؤال المطروح، مثال نظام SOQAL، وبالتالي افتقرت نظم الإجابة الآلية باللغة العربية لوجود آلية بإمكانها تقديم إجابات دقيقة وواضحة وفقاً لنمط السؤال المطروح،

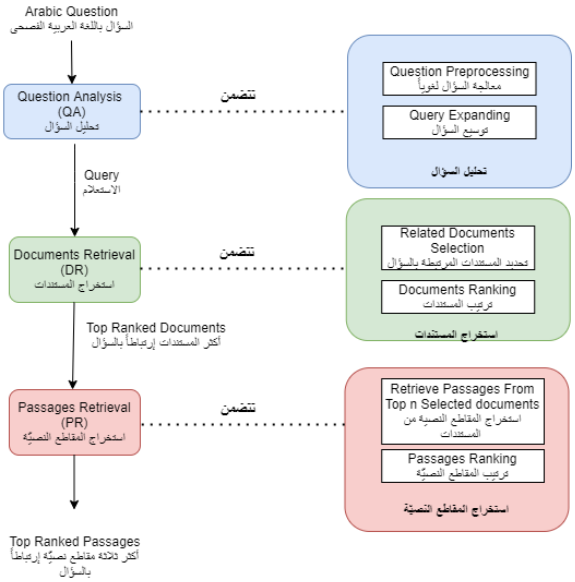
إجابة آلية يمكنه حل هذه الثغرات (قدر الإمكان)، وكخطوة أوليّة لتحقيق مثل هكذا نظام، نقترح في هذا البحث خوارزمية لتحقيق مرحلة استخراج المقاطع النصيّة في نظم الإجابة الآلية باللغة العربية.

لاستخراج المقاطع النصيّة؛ لابد من المرور أولاً بالمراحل التالية: مرحلة تحليل السؤال، ومرحلة استخراج المستندات، وبالنهاية مرحلة استخراج المقاطع النصيّة، بحيث يكون خرج كل مرحلة هو دخلاً للمرحلة التالية (انظر الشكل(1))، يتمثل دخل النظام المقترح بسؤالاً مطروحاً باللغة العربية، ويتمثل الخرج بأكثر ثلاثة مقاطع نصيّة ارتباطاً بالسؤال.

4-1- تحليل السؤال

تُطبق في هذه المرحلة مجموعة من خطوات المعالجة اللغوية على نص السؤال ليصبح بصيغة ومعتمدة على تقنيات حديثة في المعالجة الدلالية لاستخراج الإجابة.

إنطلاقاً من الثغرات الموجودة في نظم الإجابة الآلية في اللغة العربية المتاحة حالياً، نحتاج لنظام



الشكل(1) المخطط العام للنظام المقترح.

استعلام يمكن التعامل معه في المراحل اللاحقة، وهي ذاتها الخطوات المطبقة في نظام SOQAL [11].

كما قمنا في هذه المرحلة بتطبيق تقنية توسيع الاستعلام Query Expansion[10]، وذلك بهدف زيادة دقة عملية استرجاع النصوص في كل من مرحلتي استخراج المستندات واستخراج المقاطع النصيّة، حيث جرى تنفيذ ذلك بأخذ مرادف وحيد (إن وجد) لكل كلمة من كلمات السؤال وإضافته لنص السؤال، وذلك باستخدام تقنية تضمين الكلمات، واعتمدنا على نموذج تضمين كلمات مدرب مسبقاً AraVec[11] يأخذ الكلمة ويعيد أكثر كلمة متشابهة معها في السياق (تردان في سياق واحد عادةً).

---

[10] **توسيع الاستعلام:** وهي عملية يتم فيها تعزيز الاستعلام الأصلي بكلمات مرادفة أو مرتبطة بكلمات البحث، من أجل تحسين فعالية عملية استرجاع المعلومات.

[11] https://github.com/bakrianoo/aravec

يوضح الجدول (1) أمثلة لعملية توسيع السؤال باستخدام نموذج تضمين الكلماتAraVec (تم وضع الأمثلة التالية من مجموعة المعطيات المعتمدة في ARCD).

الجدول (1) مثال عن عملية توسيع الاستعلام باستخدام تقنية تضمين الكلمات.

السؤال قبل عملية التوسيع	السؤال بعد عملية التوسيع
"ماهي الكيانات الأولى في المملكة العربية السعودية؟"	"ماهي الكيانات المنظمات الأولى في المملكة العربية السعودية؟"
"من ينتقد خاشقجي في مقالاته الإخبارية؟"	"من يتنقد نقد خاشقجي في مقالاته كتاباته الإخبارية؟"

نلاحظ من المثال السابق، لم تُأخذ مرادفات كافة كلمات السؤال، ويعود ذلك لنموذج تضمين الكلمات المُستخدم والمدونات المدرب عليها (قد لا تحوي مدونات التدريب على الكلمة وبالتالي لايمكن استخراج مرادفها).

جرت عملية توسيع نص السؤال باستخدام أداة AraVec وهي عبارة عن نماذج تضمين كلمات مدربة مسبقاً، بحيث تم تجميع بيانات تدريب هذه النماذج من ثلاثة مصادر أساسية وهي: تغريدات تويتر، وصفحات الويب المتنوعة، ومقالات الويكيبيديا. يبلغ عدد الكلمات التي يمكن تمثيلها باستخدام نماذج AraVec أكثر من 3,300,000,000 كلمة [12].

4-2- استخراج المستندات

يتم في هذه المرحلة استخراج مجموعة المستندات المرتبطة بالسؤال، ودخل هذه المرحلة هو السؤال الناتج عن المرحلة السابقة (السؤال بعد المعالجة والتوسيع) على شكل استعلام للبحث عن أكثر المستندات إرتباطاً به. يتم بالبداية معالجة كافة المستندات التي تمثل مجموعة بيانات النظام، بحيث جرت خطوات المعالجة نفسها المطبّقة على نص السؤال، إضافة لخطوة إزالة كلمات الوقف[12]. قمنا باستخراج المستندات باستخدام Vector Space Model مع قياس TF-IDF، كما توضح المعادلات (1) و(2) و(3)، كما جرى تمثيل نص السؤال وفقاً لطريقة تمثيل المستندات، قمنا بقياس التشابه ما بين المستندات النصيّة ونص السؤال باستخدام قياس تشابه جيب التمام Cosine Similarity ما بين متجهات المستندات ومتجه السؤال، وجرى ترتيب المستندات وفقاً لمدى التشابه، واخترنا أكثر خمسة مستندات متشابهة مع نص السؤال كمستندات مرشحة للمرحلة التالية (مرحلة استخراج المقاطع النصيّة).

4-3- استخراج المقاطع النصيّة

تعتبر هذه المرحلة هي المرحلة الأساسية التي يرتكز عليها بحثنا، وذلك لمدى أهميتها في نظم الإجابة الآلية بشكل عام، بحيث أن عملية استخراج الإجابة الصحيحة عن السؤال المطروح مرتبطة بشكل كبير بالمقاطع النصيّة الناتجة عن هذه المرحلة. يجري في هذه المرحلة استخراج أجزاء نصيّة من المستندات الناتجة عن المرحلة السابقة (مرحلة استخراج المستندات)، بحيث نعتبر المقطع النصي هو الجزء من النص الذي ينتهي بنقطة. تُعالج المستندات وفقاً لمراحل المعالجة اللغوية المذكورة سابقاً، إضافة لمرحلة التقطيع وفقاً للنقطة "."، وينتج بذلك مجموعة من المقاطع النصيّة.

قمنا بوضع خوارزمية استخراج المقاطع النصيّة نتيجة التجريب والاختبار لعدة مقاييس ووفقاً للاختبار وتحليل النتائج جرى اعتماد النموذج النهائي لهذه المرحلة، ولنستطيع القيام بتجريب عدة مقاييس ومراقبة النتائج، لابد

[12] **كلمات الوقف** أو الكلمات المستبعدة: وهي الكلمات التي تتكرر في النصوص مثل (في، من، إلى،...) ويستحسن تجاهلها وعدم فهرستها من أجل تحسين البحث.

لنا من تحديد مقياس اختبار معين، جرى في بحثنا اعتماد قياس دقة استخراج المقاطع النصية والذي يتم حسابه وفقاً للمعادلة التالية:

$$Acc = \frac{Nb\ of\ correct\ passages}{N} \quad (6)$$

بحيث:

- Nb of correct passages تمثل عدد المقاطع النصية التي تم استخراجها بشكل صحيح، بحيث يعتبر المقطع النصي صحيحاً في حال احتوى على الإجابة الصحيحة عن السؤال المطروح في عينة الاختبار.
- N حجم عينة الاختبار مقاساً بعدد الثنائيات (سؤال، جواب).

لقياس التشابه ما بين المقطع النصي ونص السؤال، قمنا بدايةً بتجريب قياس التشابه على المستوى النحوي فقط، وذلك من خلال استخدام أحد التقنيات التقليدية والمستخدمة مسبقاً لهذا الغرض في الأبحاث المشابهة.

4-3-1- قياس التشابه نحوياً:

جرى قياس التشابه النحوي ما بين نص السؤال (بعد المعالجة) والمقاطع النصية المُستخرجة باستخدام قياس BM25 والممثل في المعادلة رقم (4)، ووفقاً للاختبار والتجريب تمكّنا من استخراج مقاطع نصيّة بشكل صحيح اعتماداً على قياس BM25 بدقة 89%، وعند تحليل النتائج تبيّن أن النظام يفشل في استخراج المقطع النصي الصحيح في الحالات التي لا يكون فيها السؤال متطابق نحوياً مع المقطع النصي كما في السؤال "ما هي مهنة خاشقجي؟"، حيث أن الإجابة الصحيحة عن هذا السؤال متضمنة في المقطع النصي "صحفي وإعلامي سعودي، رأس عدّة مناصب لعدد من الصحف في السعودية، وتقلّد منصب مستشار، كما أنّه مدير عام قناة العرب الإخبارية سابقًا."، بينما يرد النظام المقترح في اعتماد القياس السابق المقطع النصي التالي "جمال أحمد حمزة خاشقجي (13 أكتوبر 1958، المدينة المنورة – 2 أكتوبر 2018)"، والسبب في ذلك أن القياس السابق لا يأخذ الدلالة بالاعتبار، فيتم تجاهل كافة الكلمات التي قد تكون مرادفة أو متقاربة بالمعنى مع كلمات السؤال المطروح.

بناءً على النتائج السابقة قمنا بتجريب قياس آخر ينظر لدلالة الكلمات في حساب قيمة التشابه.

4-3-2- قياس التشابه دلالياً:

اعتمدنا في قياس التشابه الدلالي على تقنية تضمين الكلمات، وهي تقنية غير مستخدمة مسبقاً في مجال استخراج المقاطع النصيّة في نظم الإجابة الآلية باللغة العربية (على حد علمنا ووفقاً للدراسة المرجعية التي قمنا بإجرائها، ولكن مستخدم في النظم المشابهة في اللغة الإنكليزية). قمنا باستخدام نموذج تضمين كلمات مدرب مسبقاً AraVec (وهو نفسه المستخدم في مرحلة توسيع السؤال) بحيث يتم تمثيل كل كلمة بمتجه رقمي استناداً للسياق الواردة ضمنه في مجموعة مدونات التدريب، وتم قياس تشابه الكلمات باستخدام قياس Cosine Similarity بين هذه المتجهات. قمنا بوضع الخوارزمية التالية لحساب التشابه الدلالي بالاعتماد على نموذج تضمين الكلمات AraVec:

الدخل: جملتين نصيتين باللغة العربية الفصحى (S_1, S_2).

الخرج: قيمة التشابه بين S_1 وS_2 باستخدام نموذج تضمين الكلمات AraVec.

الخطوات:

1. تقطيع الجملة S_1 إلى كلمات ووضعها في القائمة S_{words1}.
2. تقطيع الجملة S_2 إلى كلمات ووضعها في القائمة S_{words2}.
3. تهيئة قيمة التشابه بوضع sim = 0.

4. من أجل كل كلمة w_1 في القائمة S_{words1}:

1.4.من أجل كل كلمة w_2 في القائمة S_{words2}:

1.1.4. استخراج متجه الكلمة w1 الناتج عن نموذج تضمين الكلمات ووضعه في vec_1.

2.1.4. استخراج متجه الكلمة w2 الناتج عن نموذج تضمين الكلمات ووضعه في vec_2.

3.1.4. قياس التشابه بين المتجه vec_1 وvec_2 باستخدام قياس cosine similarity بين المتجهين، وإضافة النتيجة إلى sim.

5. إعادة sim كقيمة نهائية تعبر عن مدى التشابه.

وبتجريب القياس السابق لوحده على عينة الاختبار المعتمدة، استطاع النظام أن يعيد المقاطع النصيّة الصحيحة بدقة 60%، وبتحليل النتائج تبيّن أن السبب في فشل النظام باستخراج المقاطع النصيّة الصحيحة بنسبة 40% يعود لأن القياس السابق يعتبر كلمتين متشابهتين بحسب مدى ورودهما في نفس السياق وفقاً لمدونات التدريب، مثال يعتبر القياس السابق أن كلمة "محمد" و"جمال" متقاربتين جداً وذلك لأن الكلمتين تشيران لاسم علم وتردان عادةً في نفس السياق، ولكن بالواقع هناك حالات يجري فيها السؤال عن شخص معين بينما تأتي الإجابة عن شخص آخر نتيجة هذا القياس.

إنطلاقاً من النتائج السابقة تبيّن لنا أن كلاً من القياسين له جانباً إيجابياً يفتقده القياس الآخر، فعند اعتماد خوارزمية BM25، يجري استخراج المقاطع النصية التي تشترك مع نص السؤال بالمعلومات النحوية (جذوع كلمات مشتركة)، وفي قياس التشابه باعتماد نموذج AraVec يجري استخراج المقاطع النصيّة التي تشترك مع نص السؤال في المرادفات والكلمات المتشابهة دلالياً، وبالتالي يمكن الاستفادة من القياسيين السابقين من خلال جمعهم بقياس واحد مع إضافة معامل توزين معين يعطي قيمة أكبر لأحد القياسيين على الآخر، وهذا ما قمنا بتجريبه واختباره في قياس التشابه الكلي.

4-3-3- قياس التشابه الكلي:

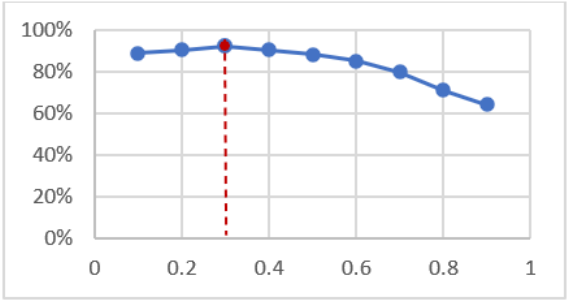
قمنا بوضع نموذج خطي لقياس التشابه النهائي ما بين نص السؤال q والمقطع النصي p، بحيث يجمع ما بين القياسين النحوي والدلالي السابقين، ويمكن تمثيله بالمعادلة التالية:

$$Sim(p,q) = \alpha\ Sim_{AraVec} + (1 - \alpha)\ Sim_{BM25} \quad (7)$$

بحيث:

- Sim(p,q) قيمة التشابه الكلي.
- SimBM25 قيمة التشابه النحوي بين المقطع النصي p والسؤال q باستخدام خوارزمية BM25.
- SimAraVec قيمة التشابه الدلالي بين المقطع النصي p والسؤال q باستخدام نموذج تضمين الكلمات AraVec.
- α معامل توزين يتم وضعه وفقاً للتجريب وقياس الدقة.

من أجل تحديد قيمة معامل التوزين α، قمنا بتجريب كافة القيم ضمن المجال [0.1,0.9] بخطوة 0.1، وقياس دقة النظام المقترح عند كل خطوة، وذلك من أجل كافة الأسئلة في عينة الاختبار المعتمدة ARCD. يبين الشكل (2) دقة استخراج المقاطع النصيّة الصحيحة وفقاً لقيمة معامل التوزين α في قياس التشابه الكلي.



الشكل (2) دقة استخراج المقاطع النصيّة الصحيحة وفقاً لقيمة معامل التوزين في قياس التشابه الكلي.

يبين الشكل السابق أن القياس المقترح يعطي أفضل نتيجة من أجل قيمة $\alpha = 0.3$ أي أن قياس التشابه على المستوى النحوي يؤثر بشكل أكبر من قياس التشابه على المستوى الدلالي في حساب التشابه الكلي.

5- النتائج

تم اختبار النظام على مجموعة البيانات ARCD، وهي مجموعة البيانات المعتمدة في نظام SOQAL[11]، استطاع النظام أن يعيد مقاطع نصيّة صحيحة بدقة 92.4% بحيث جرى حساب الدقة وفقاً للمعادلة (6). بمقارنة نظامنا المقترح مع الأنظمة المشابهة السابقة والتي قمنا بدراستها في الدراسة المرجعية، نلاحظ بأن النظام المقترح استطاع التفوق على هذه الأنظمة في بعض النقاط، فيمكن لنظامنا الإجابة عن أسئلة مطروحة ضمن مجالات متعددة وفقاً لمجموعة المعطيات التي يعتمدها؛ وبالتالي فهو غير محصور بمجال محدد، كما هو الحال في نظام ASHLK، كما أن آلية المعالجة المتّبعة في نظامنا غير مرتكزة على أنماط أسئلة محددة، بحيث يمكن استخراج مقاطع نصيّة بقطع النظر عن نمط السؤال المطروح. وبذلك يتفوق نظامنا على نظام Lemaza في قدرته على معالجة أنماط مختلفة من الأسئلة. عند مقارنة نظامنا المقترح بنظام SOQAL، حيث أن الرماز البرمجي للنظام مفتوح المصدر ومجموعة بيانات التدريب متاحة للاستخدام، جرت المقارنة وفق عدة نقاط وهي:

- قياس دقة كل من نظامنا المقترح ونظام SOQAL وفقاً لقياس الدقة الممثل بالمعادلة (6)، وباستخدام مجموعة بيانات ARCD كعينة اختبار، انظر الجدول (2)، نلاحظ بأن النظام المقترح استطاع أن يرد مقاطع نصيّة صحيحة عن الأسئلة المطروحة بدقة تصل إلى 92.4% والتي تفوق دقة نظام SOQAL بمقدار 1%.
- يعتبر نظام SOQAL نظام إجابة آلية متكامل بحيث أنه يعيد مقاطعاً نصية قصيرة (بطول 10 كلمات كحد أعظمي) كإجابات نهائية عن الأسئلة المطروحة، وبالتالي لا يحتاج SOQAL إلى مكونات ومراحل إضافية لاستخراج الإجابة النهائية، بينما يحتاج نظامنا المقترح لمرحلة إضافية تُعالج المقاطع النصية الناتجة، وتعيد إجابات دقيقة وفقاً لنمط السؤال المطروح.
- يحتاج SOQAL لمرحلة تدريب الشبكة العصبونية التي تشكل النموذج الأساسي للنظام، وعندما أجرينا مرحلة التدريب هذه وباستخدام نفس معطيات التدريب، استغرقت عملية التدريب حوالي 6 ساعات باستخدام منصة Google Colab، بينما لايوجد في نظامنا مرحلة تدريب، كما لا نحتاج معطيات تدريب، وإنما جرى وضع نموذج معين لاستخراج المقاطع النصيّة وغير مرتبط بمجموعة بيانات معينة، حيث يمكن استخدام أي مجموعة بيانات متاحة باللغة العربية.

الجدول 2 – مقارنة النظام المقترح بنظام SOQAL وفقاً لقياس الدقة.

النظام	الدقة
النظام المقترح	%92.4
SOQAL	%91.4

6- الأدوات المستخدمة

قمنا بتحقيق النظام باستخدام لغة Python، حيث جرت خطوات المعالجة اللغوية باستخدام التوابع المتاحة في مكتبة NLTK، وجرى قياس التشابه النحوي باستخدام تابع BM25 المحقق في مكتبة PyPI. قمنا بقياس التشابه الدلالي باستخدام نماذج AraVec والمحققة بلغة Python ومفتوحة المصدر.

7- الخلاصة

قدمنا في هذا البحث خوارزمية لاسترجاع المقاطع النصيّة في نظم الإجابة الآلية باللغة العربية، تعتمد على قياس التشابه النحوي والدلالي بين السؤال والمقاطع النصيّة، بحيث يقاس التشابه النحوي باستخدام خوارزمية BM25، ويقاس التشابه الدلالي باستخدام تقنية تضمين الكلمات، واعتماداً على نموذج تضمين كلمات مدرب مسبقاً. تم اختبار الخوارزمية باستخدام مجموعة بياناتARCD، وتمكّن النظام من الحصول على مقاطع نصيّة صحيحة بدقة 92.4٪. نسعى لتحسين المنهجية المقترحة من خلال استخدام تقنيات أحدث في قياس التشابه الدلالي، مثل استخدام تقنيات تضمين الكلمات السياقي ونماذج BERT اللغوية، كما نسعى لتوظيف المنهجية المقترحة في نظام إجابة آلية دقيق، وذلك من خلال تحقيق مرحلة استخراج الإجابة التي يمكن من خلالها الحصول على إجابة دقيقة وفقاً لنمط السؤال المطروح انطلاقاً من المقاطع النصيّة الناتجة.

المراجع References

[1] **Allam, Ali Mohamed Nabil, and Mohamed Hassan Haggag. "The question answering systems: A survey." International Journal of Research and Reviews in Information Sciences (IJRRIS) 2.3.(2012)**

[2] **Abdi, Asad, et al. "A question answering system in hadith using linguistic knowledge." Computer Speech & Language 60 (2020): 101023.**

[3] **Azmi, Aqil M., and Nouf A. Alshenaifi. "Lemaza: An Arabic why-question answering system." Natural Language Engineering 23.6 (2017): 877-903.**

[4] **Al-Smadi, Mohammad, et al. "Leveraging Linked Open Data to Automatically Answer Arabic Questions." IEEE Access 7 (2019): 177122-177136.**

[5] **Zhou, Mantong, et al. "Knowledge-Aided Open-Domain Question Answering." arXiv preprint arXiv:2006.05244.(2020)**

[6] **Sarrouti, Mourad, and Said Ouatik El Alaoui. "SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions." Artificial Intelligence in Medicine 102 (2020): 101767.**

[7] **Karpukhin, Vladimir, et al. "Dense Passage Retrieval for Open-Domain Question Answering." arXiv preprint arXiv:2004.04906 .(2020)**

[8] **Zuccon, Guido, et al. "Integrating and evaluating neural word embeddings in information retrieval." Proceedings of the 20th Australasian document computing symposium. 2015.**

[9] **Mitra, Bhaskar, and Nick Craswell. "Neural text embeddings for information retrieval." Proceedings of the Tenth ACM International**

Conference on Web Search and Data Mining. 2017.

[10] Li, Yang, and Tao Yang. "Word embedding for understanding natural language: a survey." Guide to Big Data Applications. Springer, Cham, 2018. 83-104.

[11] Mozannar, Hussein, et al. "Neural arabic question answering." arXiv preprint arXiv:1906.05394.**(2019**)

[12] Soliman, Abu Bakr, Kareem Eissa, and Samhaa R. El-Beltagy. "Aravec: A set of arabic word embedding models for use in arabic nlp." Procedia Computer Science 117 (2017): 256-265.

[13] Aizawa, Akiko. "An information-theoretic perspective of tf–idf measures." Information Processing & Management 39.1 (2003): 45-65.

[14] Cambridge, U. P. "Online edition (c) 2009 Cambridge UP An Introduction to Information Retrieval Christopher D." (2009).

Received	**2021/3/22**	**إيداع البحث**
Accepted for Publ.	**2021/4/29**	**قبول البحث للنشر**

مسرد المصطلحات

نموذج فضاء المتجهات	**Vector Space Model**
تـردد المصـطلح – تـردد المسـتند العكسي.	**Term Frequency Inverse Document Frequency (TF-IDF)**
الشبكات الدلالية	**Wordnets**
محركات التوصية	**Recommendation Engines**
تضمين الكلمات	**Word Embedding**
التشابه باعتماد البيان	**Graph Similarity**
الجمل الجديلة	**Cue Phrases**
التشفير ثنائي الاتجاه من محولات التمثيل	**Bidirectional Encoder from Representations Transformers (BERT)**
مجموعـة بيانـات الفهـم القرائـي العربية	**Arabic Reading Comprehension Dataset (ARCD)**
توسيع الاستعلام	**Query Expansion**
تشابه جيب التمام	**Cosine Similarity**