الجمهورية العربية السورية
المعهد العالي للعلوم التطبيقية والتكنولوجيا

ماجستير في المعلوماتية

نظام إجابة آلية باللغة العربية

Arabic Question Answering System

أُعدت هذه الأطروحة لنيل درجة الماجستير

في نظم المعطيات الكبيرة

إعداد

لانا الصّباغ

إشراف

د. أميمة الدكاك **د. ندى غنيم**

2021

Syrian Arabic Republic
Higher Institute for Applied Science and Technology
Department of Big Data

Arabic Question Answering System

Submitted in Fulfillment of the Requirements for Master's Degree
in Big Data Specialty

By

Lana Al Sabbagh

Supervised By

Dr. Nada Ghneim **Dr. Oumayma Al Dakkak**

2021

المعهد العالي للعلوم التطبيقية والتكنولوجيا

Higher Institute for Applied Sciences and Technology

المعهد العالي للعلوم التطبيقية والتكنولوجيا مؤسسة حكومية للتعليم العالي أحدثت بموجب المرسوم التشريعي رقم /24/ لعام 1983، وذلك بهدف إعداد كوادر علمية متميزة من مهندسين وباحثين للإسهام الفاعل في عملية التطوير العلمي والتنمية في الجمهورية العربية السورية.

يمنح المعهد العالي درجة الإجازة في الهندسة في الاتصالات والمعلوماتية والنظم الإلكترونية والميكاترونيكس وعلوم وهندسة المواد وهندسة الطيران. يقبل المعهد العالي لدراسة هذه الاختصاصات شريحة منتقاة من المتفوقين في الشهادة الثانوية من الفرع العلمي. يتيح المعهد العالي أيضاً برامج ماجستير أكاديمي في نظم الاتصالات وفي التحكم والروبوتيك وفي نظم المعطيات الكبيرة ونظم المعلومات ودعم القرار وفي علوم وهندسة المواد وعلوم وهندسة البصريات. وأخيراً، يمنح المعهد العالي درجة الدكتوراه في الاتصالات والمعلوماتية ونظم التحكم والفيزياء التطبيقية. تُحدث في المعهد العالي اختصاصات جديدة بحسب متطلبات سوق العمل وتوجهات البحث والتطوير المحلية والعالمية.

إلى جانب النشاط التعليمي، يمارس المعهد العالي عبر جهود أطره وفعالياته العلمية المختلفة نشاطاً حثيثاً في البحث والتطوير، إذ ينفذ مشاريع ودراسات واستشارات متنوعة لصالح الجهات العامة والخاصة في القطر، كما يتعاون مع جهات خارج القطر في بعض المشاريع البحثية والتطويرية. يسعى المعهد أيضاً، عبر دورات تدريبية نظرية وعملية متاحة للقطاعين العام والخاص وللأفراد، إلى إفادة أوسع فئة من المهتمين من إمكانيات أطره العلمية ومختبراته. واستكمالاً لدوره الرائد في مجال التعليم ونشر العلم، ينشر المعهد العالي كتباً علمية عالية المستوى من نتاج أطره، منها ما هو تدريسي يوافق المناهج في المعهد العالي ويفيد شريحة واسعة من الطلاب الجامعيين عموماً، ومنها ما هو علمي ثقافي. يتيح المعهد العالي بعضاً من منشوراته على موقعه على الشابكة، كما يتيح إمكانية الاطلاع على رسائل الماجستير والدكتوراه المنفذة في المعهد العالي وعلى بعض منشورات طلابه وأطره من المقالات العلمية.

المعهد العالي للعلوم التطبيقية والتكنولوجيا، الجمهورية العربية السورية، دمشق، ص.ب 31983

Higher Institute for Applied Sciences and Technology – HIAST

P. O. Box 31983, Damascus, Syrian Arab Republic

هاتف 00963115123819 - فاكس 00963115140761

بريد إلكتروني contact@hiast.edu.sy

موقع إلكتروني www.hiast.edu.sy

تصريح

أنا الموقع أدناه لانا محمد الصباغ معدّ أطروحة الماجستير التي تحمل العنوان: نظام إجابة آلية باللغة العربية.

أصرح بأن:

- الأعمال والنتائج المعروضة في هذه الأطروحة هي نتيجة جهودي الشخصية وبتوجيه من المشرف، وأن ما عدا ذلك من معلومات ونتائج قد نُسبت إلى مصادرها ومؤلفيها، وأُشير إلى ذلك في متن النص وفي قائمة المراجع.
- المعطيات والمعلومات المستخدمة في هذه الأطروحة جرى تحصيلها بطرائق سليمة ومشروعة ونُسبت إلى مصادرها في المواضع الملائمة.
- كلّ مكوّن من مكونات هذه الأطروحة (مقطع نصّي، صورة، مخطط، ...) مقتبس من عمل آخر جرى تمييزه بوضوح ونُسب إلى مصدره.
- الأعمال والنتائج المعروضة في هذه الأطروحة لم تُستخدم سابقًا وليست قيد الاستخدام للحصول على أي شهادة أكاديمية أخرى.

التوقيع

دمشق 11 / 8 / 2021

الشكر

أقدم كامل تقديري وشكري للأساتذة د.أميمة الدكاك ود.ندى غنيم، والمشرفيين على بحثي المتواضع، والذي اسأل الله به أن يضيف قيمة إلى هذا العلم، وأيضاً شكر موجه لإدارة المعهد العالي للعلوم التطبيقية والتكنولوجيا لتوفيرهم لنا الخدمات المطلوبة والتسهيل على الطلاب ومساعدتهم بشتى الطرق في كل الأمور التي من شأنها أن تجعل لهم فضاءً مريحاً للدراسة وطلب العلم.

Abstract

Question Answering Systems (QAS) are essential for all users, due to the increase in the amount of web information, which led to the emergence of users ’need for technologies that enable them to access direct answers to their questions. In the recent years, many researches have appeared on the Question Answering domain, but it is still a real challenge, especially for the Arabic language, because the Arabic language processing is more complicated than other languages. There are many studies on the Arabic Question Answering Systems, However, it limited for type of questions, a field of data, or the answering mechanism they used. Therefore, there must be an Arabic Question Answering Systems that cover as much as possible the gaps in the existing systems.

In this thesis, we introduce a new methodology for achieving an Arabic Question Answering System which can provide accurate and direct answers to different types of questions within several fields from a set of predefined textual documents. The proposed methodology answers questions of types: yes/no questions, factoid questions, list questions and summarization questions. The answering process goes through four phases: Question Analysis, Document Retrieval, Passage Retrieval, and Answer Extraction. Our research focuses on Passage Retrieval and Answer Extraction phases because they have a great impact on the efficiency of the system in general. In the Passage Retrieval phase, we depended on measuring the syntactic and semantic similarities between the question and the extracted passages using the newest Natural Language Processing techniques. The Answer Extraction phase was also achieved by using algorithms that extract the answer according to the type of question.

We tested our system on ARCD dataset which is used in similar previous research, it was able to answer the asked questions, and it outperformed the similar systems according to Sentence Match (SM) measure and in its ability to provide accurate and specific answers according to the type of the asked question.

الملخص

تعد نظم الإجابة الآلية مهمة جداً لكافة المستخدمين، ويعود ذلك للتزايد الكبير في كمية معلومات الوب، مما أدى لظهور حاجة المستخدمين لوجود تقنيات تمكنهم من الوصول لإجابات مباشرة عن استفساراتهم. ظهرت في الفترة الأخيرة العديد من الدراسات والأبحاث حول تقنية الإجابة الآلية، ومع ذلك، لاتزال تشكل تحدياً حقيقاً وخاصةً في مجال اللغة العربية باعتبار معالجة اللغة العربية أكثر صعوبة وتعقيداً من اللغات الأخرى. بالرغم من وجود العديد من الدراسات حول الإجابة الآلية باللغة العربية؛ إلا أنها تتقيد إما بنمط الأسئلة التي تجيب عنها، أو بمجال المعطيات التي تعالجها، أو بآلية الإجابة. لذلك، لابد من وجود نظم إجابة آلية باللغة العربية تغطي قدر الإمكان الثغرات في النظم المتواجدة حالياً.

نقدم في هذه الأطروحة منهجية جديدة في تحقيق نظام إجابة آلية باللغة العربية، بحيث تستطيع الإجابة عن أسئلة من أنماط مختلفة، وضمن مجالات عدّة، وقادرة على تقديم إجابات دقيقة ومباشرة وفقاً لنمط السؤال المطروح، وتتم عملية استخراج الإجابة من ضمن مجموعة وثائق نصيّة محددة مسبقاً. تعالج المنهجية المقترحة أسئلة من أنماطاً مختلفة وهي: أسئلة التأكيد (نعم – لا)، والأسئلة التعريفية (أين، من، ما اسم ..)، وأسئلة التعداد، وأسئلة التلخيص. يمر النظام بأربعة مراحل أساسية وهي: مرحلة تحليل السؤال، ومرحلة استخراج المستندات، ومرحلة استخراج المقاطع النصيّة، ومرحلة استخراج الإجابة. ارتكز بحثنا على إيجاد خوارزميات وطرائق جديدة في تحقيق مرحلتي استخراج المقاطع النصيّة واستخراج الإجابة، وذلك لأثرهما الكبير على كفاءة النظام بشكل عام، بحيث اعتمدنا في المنهجية المقترحة في مرحلة استخراج المقاطع النصيّة على قياس التشابه الدلالي والنحوي بين نص السؤال والمقاطع النصيّة وذلك باستخدام أحدث تقنيات معالجة اللغات الطبيعية، وكذلك تم تحقيق مرحلة استخراج الإجابة باستخدام خوارزميات تستخرج الإجابة وفقاً لنمط السؤال المطروح.

قمنا باختبار النظام المقترح على مجموعة بيانات Arabic Reading Comprehension Dataset (ARCD) والتي تم استخدامها سابقاً في أبحاث مشابهة، وجرى تقييم النظام وفقاً لمعياري Sentence Match(SM) و Macro F-Measure و استطاع بذلك التفوق على النظم المشابهة وفقاً لقياس SM وأيضاً في تمكّنه من تقديم إجابات دقيقة ومحددة وفقاً لنمط السؤال المطروح.

الفهرس

الشكر 4

Abstract 6

الملخص 7

قائمة الأشكال 12

قائمة الجداول 14

الفصل الأول: مدخل لموضوع البحث وأهدافه 16

1.تمهيد 16

2.دوافع البحث 16

3.إشكالية البحث 17

4.لمحة عن الحل المقترح 17

5.مساهمات البحث 18

6.مخطط الأطروحة 18

الفصل الثاني: دراسة نظرية 20

1.معالجة اللغات الطبيعية 20

1.1.مستوى التحليل الصرفي 20

2.1.مستوى التحليل النحوي 21

3.1.مستوى التحليل الدلالي 22

4.1.المعالجة الآلية للغة العربية 23

2.نظم الإجابة الآلية 24

1.2.مكونات نظم الإجابة الآلية 24

3.قياس تشابه النصوص 27

1.3.التشابه على المستوى النحوي 27
2.3.التشابه على المستوى الدلالي 28
4.تضمين الكلمات 30
1.4.نماذج تضمين الكلمات المستقلة عن السياق 31
2.4.نماذج تضمين الكلمات المعتمدة على السياق 31
5.الخاتمة 36
الفصل الثالث: دراسة الأعمال ذات الصلة في نظم الإجابة الآلية 38
1.معايير المقارنة..... 38
1.1.اللغة المعالجة 38
2.1.مجال المعطيات 38
3.1.أنماط الأسئلة المعالجة (الدخل) 39
4.1.شكل الخرج..... 39
5.1.آلية الاختبار والنتائج..... 40
6.1. التكلفة..... 40
2.أهم الأعمال المشابهة..... 41
1.2.الأعمال المشابهة باللغات الأجنبية 41
2.2.الأعمال المشابهة باللغة العربية 51
3.مقارنة 67
1.3.نقاط القوة في الأعمال السابقة..... 67
2.3.نقاط الضعف في الأعمال السابقة..... 67
4.الخاتمة 70
الفصل الرابع: المنهجية المقترحة 72
1.المخطط العام للمنهجية المقترحة 72
1.1.مرحلة ما قبل التشغيل Offline..... 74

2.1.مرحلة التشغيل Online 75
2.الخاتمة 86
الفصل الخامس: الاختبارات والنتائج 88
1.عينات التدريب والاختبار..... 88
1.1.عينة تدريب مصنف الأسئلة 88
2.1.عينة تدريب نموذج BERT 90
3.1.عينة الاختبار 90
2.معايير التقويم..... 91
3.نتائج اختبار النظام 92
4.اختبار الأنظمة الجزئية..... 93
1.4.اختبار مرحلة تحليل السؤال 93
2.4.اختبار مرحلة استخراج المستندات 96
3.4.اختبار مرحلة استخراج المقاطع النصيّة..... 97
4.4.اختبار مرحلة استخراج الإجابة..... 98
5.تحليل النتائج 99
الفصل السادس: الخاتمة والآفاق المستقبلية..... 102
المراجع..... 104
الملحق(1): أمثلة من عينات الاختبار 107
1.مثال عن إجابة النظام على أسئلة التأكيد..... 107
2.مثال عن إجابة النظام على الأسئلة التعريفية..... 108
3.مثال عن إجابة النظام على أسئلة التعداد 109
4.مثال عن إجابة النظام على أسئلة التلخيص..... 110
الملحق(2): النشرة المرتبطة بالبحث 111

قائمة الأشكال

الشكل 1- مثال عن شجرة التحليل لجملة ما..... 22
الشكل 2 – الشكل العام لأنظمة الإجابة الآلية..... 25
الشكل 3 – بعض مقاييس التشابه على المستوى المعجمي 28
الشكل 4 – بعض مقاييس التشابه الدلالي القائم على المعرفة 29
الشكل 5 – بعض مقاييس التشابه الدلالي القائم على المدونات 30
الشكل 6 – البنية العامة لنموذج Bert في مرحلة التدريب..... 33
الشكل 7 – تمثيل دخل نموذج Bert. 34
الشكل 8 – تخصيص نموذج Bert في الإجابة الآلية. 35
الشكل 9 – الشكل العام لنظام DrQA..... 43
الشكل 10 – البنية العامة لنظام SemBioNLQA. 45
الشكل 11 – منهجية توليد مجموعة الإجابات المقترحة في نظام EfficentQA..... 48
الشكل 12- منهجية ترميز نص السؤال في نظام EfficentQA. 49
الشكل 13 – منهجية ترميز الإجابة في نظام EfficentQA. 50
الشكل 14 – خوارزمية حساب entailment relation في نظام EWQA..... 52
الشكل 15 – البنية العامة لنظام Lemaza..... 54
الشكل 16 – البنية العامة لنظام LOD..... 56
الشكل 17 – البنية العامة لنظام Hybrid QAS. 58
الشكل 18- نتائج نظام Hybrid QAS. 59
الشكل 19- البنية العامة لنظام SOQAL. 62
الشكل 20 – البنية العامة لنظام ASHLK..... 64

الشكل 21 – البنية العامة لنموذج AraELECTRA. 66
الشكل22 – البنية العامة للنظام المقترح. 73
الشكل 23 – أمثلة عن توسيع السؤال باستخدام نموذج AraVec. 77
الشكل 24 – خوارزمية حساب التشابه باعتماد نموذج تضمين الكلمات AraVec. 80
الشكل 25 – خوارزمية الإجابة على أسئلة التأكيد..... 82
الشكل 26 – خوارزمية الإجابة على الأسئلة التعريفية..... 84
الشكل 27 – استبيان جمع عينة تدريب مصنف الأسئلة. 89
الشكل 28 – توّزع عينة تدريب مصنف الأسئلة وفقاً للأنماط المعتمدة. 89
الشكل 29 – البنية العامة لعينة اختبار النظام..... 91
الشكل 30– قياس دقة مرحلة تحليل السؤال وفقاً لخطوات المعالجة اللغوية المعتمدة..... 94
الشكل 31 – قياس دقة النظام وفقاً لنوع مصنف السؤال. 95
الشكل 32 – قياس دقة النظام وفقاً لآلية توسيع السؤال المعتمدة..... 96
الشكل 33 – قياس دقة النظام وفقاً للمعالجة اللغوية المعتمدة في مرحلة استخراج المستندات. 97
الشكل 34 – قياس دقة النظام وفقاً للآلية المتبعة في استخراج المقاطع النصيّة. 98
الشكل 35 – مثال عن إجابة النظام المقترح على أسئلة التأكيد. 107
الشكل 36 – مثال عن إجابة النظام المقترح على الأسئلة التعريفية..... 108
الشكل 37 – مثال عن إجابة النظام المقترح على أسئلة التعداد. 109
الشكل 38 – مثال عن إجابة النظام المقترح على أسئلة التلخيص. 110
الشكل 39 – مرفق قبول الورقة البحثية العربية. 111

قائمة الجداول

الجدول 1 – مقارنة بين النماذج المستقلة عن السياق والنماذج المعتمدة على السياق في عملية تضمين الكلمات [11]. 36
الجدول 2 –بعض الأنماط الرئيسية للأسئلة. 39
الجدول 3 – نتائج بحث SimBioNLQA 46
الجدول 4 – مقارنة للدراسات السابقة المشابهة. 69
الجدول 5 – أنماط الأسئلة التي يعالجها البحث. 76
الجدول 6 – توزع الأنماط المختلفة للأسئلة ضمن عينة الاختبار..... 91
الجدول 7 – نتائج اختبار النظام..... 92
الجدول 8 – مقارنة النظام المقترح بنظام SOQAL..... 92
الجدول 9 – مقارنة إجابات النظام المقترح بإجابات نظام SOQAL 93
الجدول 10 – دقة خوارزميات استخراج الإجابات. 98

الفصل الأول: مدخل لموضوع البحث وأهدافه

1.تمهيد

لم تعد عملية البحث باستخدام شبكات الوب عملية نادرة الحدوث، فنحن اليوم نقوم وبشكل مستمر بالبحث والاستفسار عن أبسط الأمور التي نواجهها. عادةً ما يحتاج المستخدم الوصول لإجابة دقيقة ومختصرة عن الاستفسارات التي يقوم بطرحها، ولكن أقصى ما يمكن لمحركات البحث تقديمه هو مجموعة من المستندات والنصوص الأكثر ارتباطاً باستفسار المستخدم، وحديثاً أصبح بإمكان محركات البحث الإشارة إلى المقطع النصي Passage الأكثر ارتباطاً باستفسار المستخدم، ويبقى للمستخدم مهمة البحث ضمن هذه المستندات والمقاطع النصيّة لإيجاد الإجابة الواضحة والمختصرة لاستفساره[1]. من هذه الغاية انطلقت نظم الإجابة الآلية والتي لاقت اهتماماً كبيراً ضمن مجال معالجة اللغات الطبيعية، وطُرح في الفترة الأخيرة العديد من الدراسات التي عُنيت بهذا المجال ومنها ما حقق نتائج جيدة جداً. ولما كان هذا التطور منتشراً بكثرة للغات الأجنبية ومحدوداً للغتنا العربية، ومع انتشار الثقافة التقنية العربية؛ كان لابد من ظهور تقنيات داعمة للغة العربية، ومن هنا تمهد الطريق أمام بحثنا لتطوير نظام إجابة على أسئلة باللغة العربية.

2.دوافع البحث

على الرغم من وجود العديد من الأبحاث حول تطوير نظم إجابة آلية باللغة العربية إلا أنها لم تكن شاملة، بعض هذه الأبحاث اختصت بالإجابة عن أنماط معينة من الأسئلة (مثال أسئلة من نوع "لماذا" أو الأسئلة الزمنية)[2]، ومنها ما اهتم بالإجابة على الأسئلة المطروحة ضمن مجال معين (مثال الإجابة عن الأسئلة الإسلامية فقط)[3]، وهناك بعض النظم اهتمت باستخراج الإجابات من مجموعة معطيات محددة البنية (مثال استخراج الإجابات من أنطولوجيا)[4]، ويوجد نظم إجابة آلية باللغة العربية مفتوحة المجال وغير محددة بنمط أسئلة معينة، ولكنها تقتصر على رد مقطعاً نصياً يحوي على الإجابة بدلاً من الإجابة الدقيقة وفقاً لنمط السؤال المطروح (لا يوجد معالجة لنمط السؤال)[5]، وبذلك لم نستطع الحصول لحد الآن على نظم إجابة آلية باللغة

العربية شاملة، بحيث تستطيع الإجابة الدقيقة عن أنماط متعددة من الأسئلة ضمن مجالات مختلفة ومن مجموعة معطيات غير محددة ببنية معينة (مجموعة نصوص).

3.إشكالية البحث

يرتكز هذا البحث على إيجاد طريقة جديدة في الإجابة الدقيقة عن الأسئلة المطروحة باللغة العربية الفصحى ضمن مجالات مفتوحة، وغير مقيدة بنمط سؤال محدد وذلك من مجموعة وثائق نصيّة، وعلى رغم من وجود دراسات سابقة ضمن نفس المجال؛ إلا أنها كانت مقتصرة على معالجة الأسئلة ضمن شروط معينة (كنمط السؤال أو مجال الإجابة)، وإن كانت شاملة فلم تكن معتمدة على تقنيات معالجة لغوية حديثة وفعّالة مما ينعكس على الدقة النهائية للنظام.

ندرس في هذا البحث إمكانية إيجاد منهجية جديدة لتحقيق عملية الإجابة الآلية باللغة العربية بحيث تكون شاملة قدر الإمكان، وتغطي النقص في الدراسات السابقة المشابهة، وتعتمد على تقنيات حديثة تزيد من دقة الحصول على الإجابة الصحيحة، كما نعمل على تقويم مدى جودة الحل المقترح ومقارنته بالطرق الأخرى.

4.لمحة عن الحل المقترح

تتكون نظم الإجابة الآلية عادةً من أربع مكوّنات رئيسية وهي على التتالي مكوّن **معالجة الاستعلام** (مكوّن معالجة السؤال)، ومكوّن **استخراج المستندات**، ومكوّن **استخراج المقاطع النصيّة** ومكوّن **استخراج الإجابة**. بحيث يتم في مكوّن معالجة السؤال إجراء خطوات المعالجة اللغوية لنص السؤال لتتم مطابقته لاحقاً مع المستندات النصية التي تشكل مجموعة بيانات النظام وذلك ضمن مكوّن استخراج المستندات، والذي يعيد أكثر المستندات ارتباطاً بنص السؤال (يمكن تحديد أفضل n مستنداً)، ليتم بعدها استخراج المقاطع النصيّة الأكثر ارتباطاً بالسؤال، وذلك ضمن مكوّن استخراج المقاطع النصيّة، وأخيراً يتم استخراج الإجابة الدقيقة والمحددة للسؤال من ضمن المقاطع النصيّة الناتجة، وذلك ضمن مكوّن استخراج الإجابة[6].

يعتبر مكوّن استخراج المقاطع النصيّة من أكثر المكونات أهمية في نظم الإجابة الآلية، ويعود ذلك لارتباط صحة الإجابة المستخرجة من المقاطع النصيّة بصحة المقاطع النصيّة الناتجة عن هذا المكون، وتنعكس دقة هذا المكوّن بشكل كبير على الدقة الإجمالية في النظام، ولهذا السبب ارتكز بحثنا على إيجاد طريقة فعّالة في تحقيق مكوّن استخراج المقاطع النصية بحيث تم ذلك باستخدام تقنيات حديثة في مجال معالجة اللغات الطبيعية تعتمد على قياس التشابه النحوي والدلالي بين نص السؤال والمقاطع النصية المستخرجة لاستخلاص الإجابة الدقيقة للسؤال. كما عملنا على زيادة دقة النظام من خلال استخدام تقنيات معالجة لغوية حديثة في مكوّني معالجة السؤال واستخراج الإجابة.

5.مساهمات البحث

يمكن تلخيص المساهمات الأساسية لبحثنا في النقاط التالية:

- طرح خوارزمية جديدة في تحقيق مكوّن استخراج المقاطع النصيّة في نظام الإجابة الآلية بحيث تأخذ بالاعتبار التشابه النحوي والدلالي ما بين المقطع النصي ونص السؤال، بحيث تم تطبيق الخوارزمية على مجموعة البيانات المستخدمة ذاتها في أبحاث مشابهة سابقة [5]، و أثبتت تفوقها على الدراسات السابقة في هذا المجال.
- تطبيق تقنية لغوية حديثة في معالجة السؤال تعتمد على توسيع نص السؤال بإضافة مرادفات كلماته مما يزيد من دقة النظام، اعتمدنا في تحقيق عملية التوسيع هذه على مفهوم تضمين الكلمات، والذي سنتطرق لشرحه لاحقاً ضمن هذه الأطروحة، بحيث بيّنت القياسات بأن استخدام هذا المفهوم قد زاد من الدقة الإجمالية للنظام وتفوق على أداء التقنيات المشابهة المستخدمة لتحقيق مرحلة توسيع السؤال.
- تحقيق مكوّن استخراج الإجابة عن طريق تطبيق منهجية جديدة مستوحاة من النظم المشابهة الحديثة ولكن على اللغة الإنكليزية [7]، ولم يسبق تطبيقها على اللغة العربية في الدراسات السابقة مما دفعنا لتطبيقها في بحثنا وأثبتت كفاءتها في الدقة الإجمالية للنظام.

6.مخطط الأطروحة

قمنا في هذا الفصل بتحديد مشكلة البحث، وتقديم لمحة عن الحل المقترح والمساهمات الأساسية التي يقدمها. سنتابع في تتمة الأطروحة بدراسة نظرية لبعض المفاهيم الأساسية في مجال معالجة اللغات الطبيعية بشكل عام، ونظم الإجابة الآلية خاصةً، وذلك ضمن **الفصل الثاني**، كما سنستعرض ضمن **الفصل الثالث** أهم الأبحاث والدراسات المشابهة الحديثة ضمن نفس المجال والمقارنة بينها؛ مع توضيح القيمة المضافة التي سيقدمها بحثنا إضافةً عن هذه الأبحاث. يعرض **الفصل الرابع** المكوّنات الأساسية للنظام مع شرح تفصيلي للمراحل التي يمر بها كل مكوّن، ليتم في **الفصل الخامس** اختبار النظام وفقاً لعدة معايير ومقارنة أدائه بأداء الأنظمة المشابهة السابقة. ونختم بحثنا في **الفصل السادس** بعرض الآفاق المستقبلية لهذا العمل.

الفصل الثاني: دراسة نظرية

1.معالجة اللغات الطبيعية

يأمل العلماء في مجال علوم الحاسوب بتطوير برامج حاسوبية قادرة على التفاعل مع المستخدمين باستخدام لغتهم المحكية، دون الحاجة إلى أوامر مكتوبة بلغات برمجية. إن تحقيق التواصل عن طريق اللغات الطبيعة يعزز اندماج المستخدمين مع الآلة والتفاعل معها بشكل طبيعي. ومعالجة اللغات الطبيعية هي واحدة من فروع الذكاء الصنعي وعلم اللسانيات، إذ أنها نتاج التفاعل بين لغة الحاسوب ولغة البشر وهي تعزز مشاركة البيانات، وقد حصدت نتائج جيدة في تشاركية البيانات فأصبح من السهل التعامل مع الوثائق المكتوبة بلغات طبيعية وفهمها من قبل الحاسوب وفهم التسجيلات والأوامر الصوتية [8].

على سبيل المثال، يُعدّ التحليل الصرفي الوسيلةَ الأساسية التي ترتكز عليها معظم الأنظمة الذكية التي تعالج اللغات الطبيعية، كأنظمة استعادة المعلومات، وتصنيف النصوص، والترجمة الآلية وغيرها، وتتوفر عدة مصادر مفتوحة تدعم معالجة اللغات الطبيعية.

تمر معالجة اللغة الطبيعية بعدة مستويات تحليل، وتختلف مراحل المعالجة ضمن كل مستوى، أهمها مستوى التحليل الصرفي ومستوى التحليل النحوي ومستوى التحليل الدلالي.

1.1.مستوى التحليل الصرفي

وهو الجزء الذي يهتم في معرفة نوع الكلمات واحتوائها على الضمائر وغيرها من المعلومات الصرفية. يتضمن التحليل الصرفي إيجاد جذع أو جذر[1] الكلمة وتجريدها من الزوائد (السوابق واللواحق) مثال كلمة "سنؤتيكهم" جذعها "نؤتي" وجذرها "آتى"، يستفاد من التحليل الصرفي في رد الكلمات إلى جذوعها وبالتالي تقليص حجم الكلمات المعالجة، بالإضافة إلى استنباط بعض الخصائص من تحليل الكلمة إلى مركباتها، ومن هذه الخصائص دلالة الكلمة هل تدل على الجمع أو التثنية أو الإفراد، وزمن

[1] **الجذع**: هو الكلمة بعد حذف السوابق واللواحق منها، أما **الجذر**، فهو أبسط مكون في اللغة (لايمكن تحليله لأجزاء أبسط)

الأفعال، والتذكير والتأنيث للأسماء، وأحرف الجر المتصلة بالكلمة والتي تحدد حركتها الإعرابية، والضمائر المتصلة والتي تفيد في ربط كلمات ذات السياق القريب. من هذه الخصائص ما يلزم أثناء المعالجة على المستوى الصرفي ومنه ما يلزم لمراحل متقدمة أثناء التحليل النحوي أو تحديد صنف الكلام وحتى المستوى الدلالي.

لا يقتصر التحليل الصرفي على تجزئة الكلمة إلى مركباتها فحسب، وإنما يمكن أن يتعدى ذلك لإيجاد جذور الكلمات. يستفاد من إيجاد الجذور بتقليص حجم الكلمات المعالجة بشكل أكبر من التقليص عند استخدام الرد إلى الجذع فقط. تعتمد بعض تطبيقات معالجة اللغات على رد الكلمات إلى جذعها وبعضها يعتمد على الرد إلى الجذور وذلك حسب الدقة الأعلى عند تطبيق كل منهما. ومن الأدوات البرمجية التي تعمل كمحللات صرفية للغة العربية khoja Stemmer[1] و ISRI Stemmer[2] والخليل الصرفي[3] وغيرها.

2.1.مستوى التحليل النحوي

وهو الجزء الذي يهتم بعلاقة الكلمات بعضها مع بعض، وهيكيلية الجملة، وغيرها من المعلومات النحوية، ويعتمد على مستوى التحليل الصرفي. هناك عدة مسائل جزئية تندرج تحت مستوى التحليل الصرفي وأهمها:

1.2.1.تحديد أنماط الكلمات pos-tagging (Part Of Speech Tagging)

يمكن تصنيف الكلمات في أي لغة إلى ثلاثة أصناف (فعل ، اسم ، حرف)، ولكن في اللغة العربية يعتمد هذا التصنيف على موضع الكلمة ضمن السياق، فكلمة "لعب" يمكن أن تصنف على أنها فعل ماض "لَعِبَ" عند القول:"لَعِبَ الطفل بالكرة"، ومن الممكن أن تصنف كاسم "لِعْبٌ" عند القول:"لِعبٌ جميلٌ"، كما أن هذه المرحلة من المعالجة تحل الغموض الذي يحصل في المستوى الصرفي، فحينما تدخل كلمة "كوارث" في مرحلة الصرف ينتج عنها احتمالان، إما ("ك "حرف جر وتشبيه + "وارث" اسم مفرد)، أو ("كوارث" اسم جمع مفرده كارثة)، ولكن عند المرور بمرحلة تحديد صنف الكلام يستطيع المصنف تحديد الاحتمال الصحيح للكلمة حسب سياقها. من أشهر البرمجيات التي تعمل كمحدد صنف كلام للغة العربية Stanford Parser[4]، ويقوم بإيجاد صنف الكلام اعتماداً على نموذج مدرب مسبقاً على مجموعة كبيرة من الأمثلة باللغة العربية، ويمكن

---

[1] https://github.com/motazsaad/khoja-stemmer-command-line

[2] ISRI: Information Science Research Institute, https://www.nltk.org/_modules/nltk/stem/isri.html

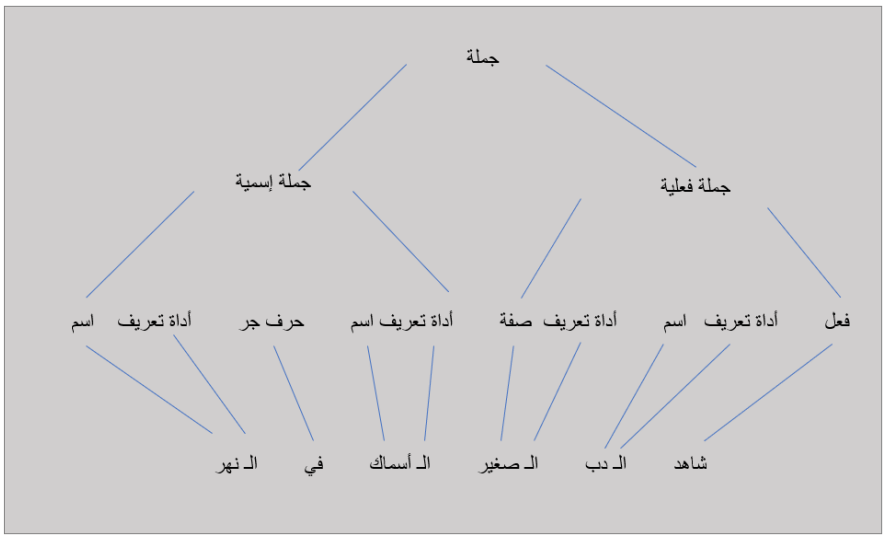
[3] https://ossl.alecso.org/affich_oso_details.php?id=57

[4] http://nlp.stanford.edu:8080/parser/

تضمين المحلل المدرب بشكل كامل في أي نظام لأنه مفتوح المصدر كما يمكن تصدير خرجه لمراحل متقدمة في المعالجة، تؤمّن هذه الأداة أيضاً نماذج مدربة لكل من التحليل النحوي للجملة العربية، والتحليل الصرفي وتجزئة النص العربي إلى كلمات.

2.2.1.التحليل النحوي السطحي

تسمح القواعد النحوية للغة بإنشاء جملٍ ضمن هذه اللغة. والتحليل النحوي بشكل عام هو عملية إيجاد شجرة تحليل متوافقة مع القواعد النحوية للغة؛ أي إيجاد مجموعة القواعد النحوية وتسلسلها الذي أدى إلى إنشاء الجملة. يمكن لشجرة التحليل أن تعطينا معلومات حول أنماط الكلمات POS، كما يمكن أن تعطينا أيضًا مجموعة الكلمات المرتبطة مع بعضها لتشكل عبارات، وكذلك العلاقة بين هذه العبارات كما في الشكل(1).



الشكل 1 – مثال عن شجرة التحليل لجملة ما.

التحليل النحوي هو عملية الحصول على جزء من شجرة التحليل. بحيث يمكن القول إن عملية إيجاد أصناف الكلمات هي عملية الحصول على الطبقة الأخيرة من شجرة التحليل. بينما التحليل النحوي السطحي هو إيجاد الكلمات التي يمكن تجميعها معًا لتكوين عبارات أي الحصول على آخر طبقتين من شجرة التحليل (مثال تقسيم الجملة إلى عبارات اسمية NP[1] وفعلية[2]VP).

3.1.مستوى التحليل الدلالي

يهتم التحليل الدلالي بدراسة المعنى الدقيق للكلمات والجمل النصية، ويتم التحليل الدلالي على جزأين أساسيين وهما:

---

[1] NP: Noun Phrase

[2] VP: Verb Phrase

دراسة معنى الكلمة الفردية :ويتم فيه الكشف عن معنى الكلمات بشكل فردي، بقطع النظر عن السياق الواردة ضمنه، ويطلق عليه أيضاً الدلالة المعجمية ، ودراسة تركيب الكلمات الفردية، ويتم فيه استخلاص المعنى العام للجملة اعتماداً على معاني كلماتها وترابط هذه المعاني فيما بينها.

إن المهمة الأساسية في التحليل الدلالي ليس فقط الكشف عن معنى الجملة وإنما كشف المعنى الصحيح لها، فمثلاً في الجملة "أخذ الولد يلعب" كلمة "أخذ" ممكن أن تأخذ معنى "بدأ باللعب"، أو "تناول الشيء أو أخذه"، ولكن ما يحدد المعنى الصحيح هو سياق الجملة وهذه هي مهمة المحلل الدلالي.

4.1.المعالجة الآلية للغة العربية

اللغة العربية هي سادس أكثر اللغات المحكية في العالم واللغة الأم لثلاثمئة مليون شخص، واللغة الرسمية لاثنين وعشرين دولة بالإضافة لكونها لغة الإسلام. تكتب من اليمين إلى اليسار وتتألف من 28 حرفاً. أغلب الكلمات العربية مشتقة صرفياً من أكثر من جذر وأغلب هذه الجذور ثلاثية تتكون من ثلاث حروف ساكنة.

1.4.1.نقاط ضعف معالجة اللغة العربية

تختلف اللغة العربية عن اللغات الأخرى فهي تمتلك سمات تميزها عن سائر اللغات. لذلك تختلف معالجة اللغة العربية عن المعالجة المتبعة في اللغات الأجنبية، وعلى الرغم من وجود مصادر وأنظمة تعالج اللغة العربية إلا أنها في حالتها البدائية إلى الآن، ولم تصل إلى مستوى الأنظمة التي تعالج اللغات الأخرى. نوضح فيما يلي بعض الخصائص التي تبطئ من عملية تطور الأنظمة التي تعالج اللغة العربية:

- غياب التشكيل من معظم النصوص العربية يشكل حالة غموض أثناء التحليل الصرفي وإيجاد شجرة التحليل القواعدي الصحيح للجملة.
- اختلاف شكل المحرف المكتوب تبعاً لموضعه ضمن الكلمة.
- لا يوجد أحرف كبيرة (Capital Letter) مما يصعب عملية التحقق من كون الكلمة اسم شخص أو مدينة أو مكان وبالتالي صعوبة تحديد الكيان الذي تنتمي إليه الكلمة (Named Entity).
- قلة المعاجم الإلكترونية للغة العربية المصنفة حسب المجال.

2.4.1.نقاط القوة لتطوير معالجة اللغة العربية

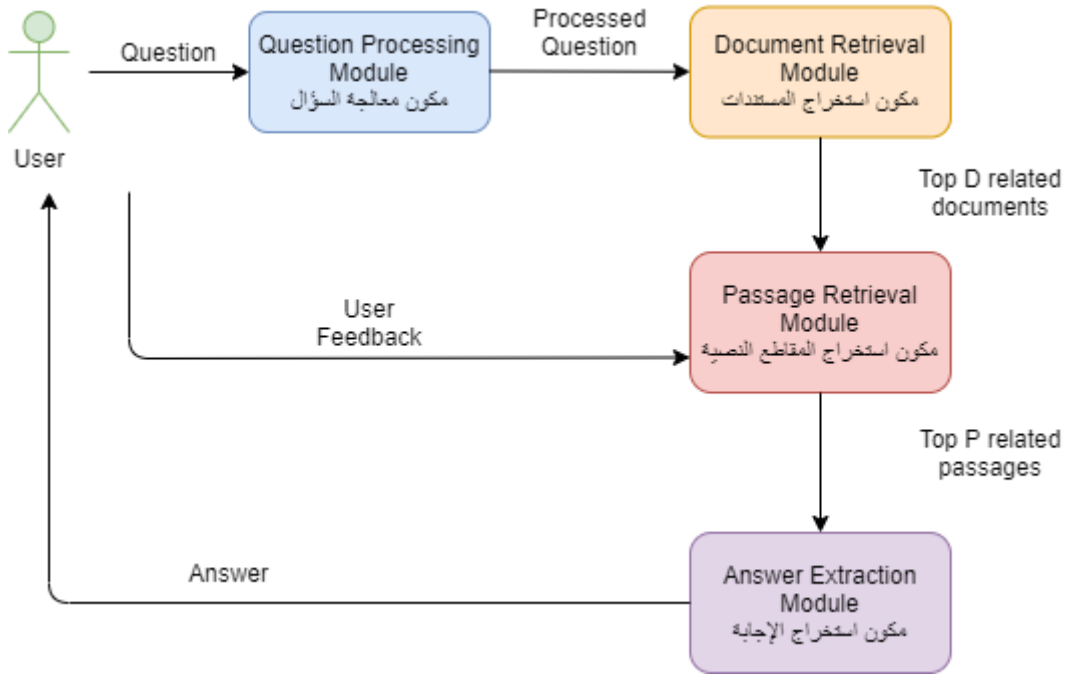
أصبح من الضروري تطوير النظم الآلية الداعمة للغة العربية، وذلك نظراً لتزايد المحتوى العربي الرقمي بالإضافة لزيادة المستخدمين للوسائل والتقنيات الحديثة التي تغني هذا المحتوى المتوفر على الانترنت؛ فلابد من إنشاء أنظمة تعالج اللغة العربية ليسهل التفاعل مع المستخدم العربي وتوقع رغباته واحتياجاته المستقبلية.

2.نظم الإجابة الآلية

تزايدت في الآونة الأخيرة المعلومات المتوفرة على الإنترنت بشكل كبير، ويقوم أكثر المستخدمين بطرح سؤال محدد بهدف البحث عن جواب محدد له، ويفضّلون الحصول على جواب مختصر وملم بالمعلومة المراد اكتشافها، ومكتوب بلغتهم المحكية. تؤمن نظم الإجابة الآلية Question Answering Systems(QAS) القدرة على الإجابة عن الأسئلة المكتوبة باللغة المحكية، وذلك من خلال استخلاص الجواب من الوثائق المخزنة التي تحوي معلومات ضمن مجال السؤال المطروح. يهتم مستخدمو نظم الإجابة عن الأسئلة بالحصول على جواب مختصر ومفهوم وصحيح، والذي يمكن أن يكون كلمة، أو جملة، أو نص، أو صوت، أو فيديو، أو صورة. تلعب نظم الإجابة الآلية دوراً هاماً في المجالات التقنية الحالية، وتتجه بشكل مباشر للمستخدمين الذين يبحثون عن إجابات دقيقة وعميقة [1]. يمكن تصنيف نظم الإجابة الآلية إلى صفين؛ إما محددة المجال (closed domain or restricted domain)، و تعتمد هذه النظم على معرفة ضمن مجال محدد ولا تجيب إلا على الأسئلة من ضمن مجال المعرفة التي يستند عليها، ولكن بالمقابل يعطي إجابة دقيقة وعميقة، أما النوع الآخر فهو النظم المفتوحة المجال (open domain)، وهي النظم التي تشمل الإجابة عن الأسئلة ضمن أي مجال كانت، وتعتمد على معرفة معممة وهنا ستكون الإجابة عامة وأقل عمقاً. يكون البحث ضمن نظم الإجابة الآلية بأحد الطريقين؛ إما اعتماداً على كلمات مفتاحية (keywords based)، أو اعتماداً على المعنى (semantic). عند الاعتماد على الكلمات المفتاحية يمكن أن يكون للكلمة أكثر من معنى، وهنا قد نحصل على إجابات خاطئة، لذا نضطر لفهم معنى كلمات السؤال المطروح للحصول على إجابة أكثر صحة. إذن، لابد من الاعتماد على المعنى، ويمكن أن نحصل بذلك على نتائج عالية الدقة حيث تستخدم هذه الطريقة عملية البحث ضمن مصادر معلومات (مثال ontology) لتستخلص الإجابة الصحيحة منها.

1.2.مكونات نظم الإجابة الآلية

تتألف نظم الإجابة الآلية بشكل عام من أربع مكونات أساسية، وهي على التتالي مكوّن معالجة السؤال، ومكوّن استخراج المستندات، ومكوّن استخراج المقاطع النصية، ومكوّن استخراج الإجابة [7]. يمكن توضيح النموذج العام لنظم الإجابة الآلية بالشكل(2).



الشكل 2 – الشكل العام لأنظمة الإجابة الآلية

1.1.2.مكوّن معالجة السؤال

يتمثل دخل هذا المكون بالسؤال المطروح من قبل المستخدم باللغة الطبيعية، ويمكن أن يشمل مراحل المعالجة التالية:

- تصنيف السؤال، ويتم ضمن هذه المرحلة الكشف عن نوع السؤال عبر تصنيفه لأحد أصناف رئيسية معرفة مسبقاً (سؤال نعم|لا، سؤال تعريفي، سؤال تلخيص...).
- تطبيق المعالجة اللغوية على السؤال، ومنها (تقطيع الاستعلام، وحذف كلمات الوقف، واستبعاد الكلمات والرموز غير المرتبطة بأبجدية اللغة، وتجذيع كلمات الاستعلام وغيرها).
- توسيع السؤال، ويعني إضافة كلمات تغني السؤال وتزيد من دقة نتائج النظام، ويمكن في هذه المرحلة توسيع الاستعلام بأخذ المرادف الأول لكل كلمة من كلماته، وهناك عدة تقنيات أخرى يمكن تطبيقها في هذه المرحلة.

ويكون بذلك خرج المكوّن هو السؤال بعد تطبيق عمليات المعالجة عليه، ليدخل لمكوّن استخراج المستندات، إضافة لنوع السؤال الذي يمثل دخلاً لمكوّن استخراج الإجابة.

2.1.2.مكوّن استخراج المستندات

دخل هذا المكوّن هو خرج المرحلة السابقة؛ أي السؤال بعد عمليات المعالجة المطبّقة في المرحلة السابقة، ويمكن أن يشمل على مراحل المعالجة التالية:

- معالجة المستندات التي تشكل مجموعة بيانات النظام وذلك وفقاً لخطوات المعالجة نفسها المطبّقة على سؤال المستخدم والمذكورة في فقرة (مكوّن معالجة السؤال).
- استخراج D مستنداً الأكثر ارتباطاً باستعلام المستخدم.
- إعادة ترتيب المستندات الناتجة عن المرحلة السابقة وفقاً لمعيار ترتيب محدد (قياس تشابه محتوى المستند مع نص السؤال المطروح).

ويكون بذلك خرج المكون هو D مستنداً الأكثر ارتباطاً بسؤال المستخدم.

3.1.2.مكوّن استخراج المقاطع النصية

دخل هذا المكوّن هو المستندات الناتجة عن المرحلة السابقة، ويمكن أن يشمل على مراحل المعالجة التالية:

- تقطيع المستندات إلى عدة مقاطع وفقاً لمعيار تقطيع معين، بحيث يمكن اعتبار المقطع هو الجملة الواحدة أو عدة جمل منتهية بنقطة، ويتم اعتماد ذلك وفقاً لدقة نتائج النظام. ويمكن أن يتم التقطيع باستخدام إحدى أدوات المعالجة اللغوية المتاحة أو تطوير أداة خاصة وفقاً لما يلزم.
- استخراج P مقطعاً نصياً الأكثر ارتباطاً بسؤال المستخدم.
- إعادة ترتيب المقاطع النصيّة الناتجة عن المرحلة السابقة وفقاً لمعيار محدد (قياس تشابه محتوى المقطع النصي مع نص السؤال المطروح).

ويكون بذلك خرج المكوّن هو P مقطعاً الأكثر ارتباطاً بسؤال المستخدم.

4.1.2.مكوّن استخراج الإجابة

يتمثل دخل المكوّن بالمقاطع النصية الناتجة عن المرحلة السابقة، ويتم استخراج الإجابة من المقاطع النصيّة وفقاً لنوع السؤال الناتج عن مكوّن معالجة السؤال. ويمكن تحقيق ذلك بكتابة مجموعة من القواعد الخاصة باستخراج الإجابة لكل نمط من أنماط الأسئلة الممكنة. مثال: يمكن استخراج إجابة الأسئلة من نوع المكاني عن طريق استخراج الكلمات من المقاطع النصيّة والتي يندرج تصنيفها كمكان، ويمكن القيام بذلك بالاستعانة بتقنية التعرف على الكيانات المسماة Named Entity Recognition (NER)[1].

[1] يعتبر **الكيان المسمى** كائنًا حقيقيًا مثل الأشخاص، والمواقع، والمؤسسات، والمنتجات وما إلى ذلك، والتي يمكن الإشارة إليه باسم علم. يمكن أن تكون هذه الكيانات مجردة أو لها وجود مادي. من الأمثلة على الكيانات المسماة باراك أوباما، مدينة نيويورك، فولكس فاجن جولف، أو أي شيء آخر يمكن تسميته. يمكن ببساطة النظر للكيانات المسماة على أنها مثال لكيان (على سبيل المثال، مدينة نيويورك هي مثال لمدينة).

ويكون بذلك خرج المكوّن هو الجواب المقتطع من المقاطع النصية الأكثر ارتباطاً بسؤال المستخدم.

يعتبر مكونا استخراج المستندات واستخراج المقاطع النصية من المكونات الهامة جداً في نظم الإجابة الآلية، ويعود ذلك لأثرها الكبير على النتائج النهائية للنظام، فترتبط دقة الإجابة بدقة الحصول على المقطع النصي الصحيح المستخرجة منه. تعتمد نظم الإجابة الآلية في تحقيق مثل هذه المكونات على قياس التشابه بين النص المستخرج ونص السؤال المطروح. سندرس في الفقرة التالية الآليات المختلفة لقياس التشابه بين النصوص والمعتمدة في بعض نظم الإجابات الآلية.

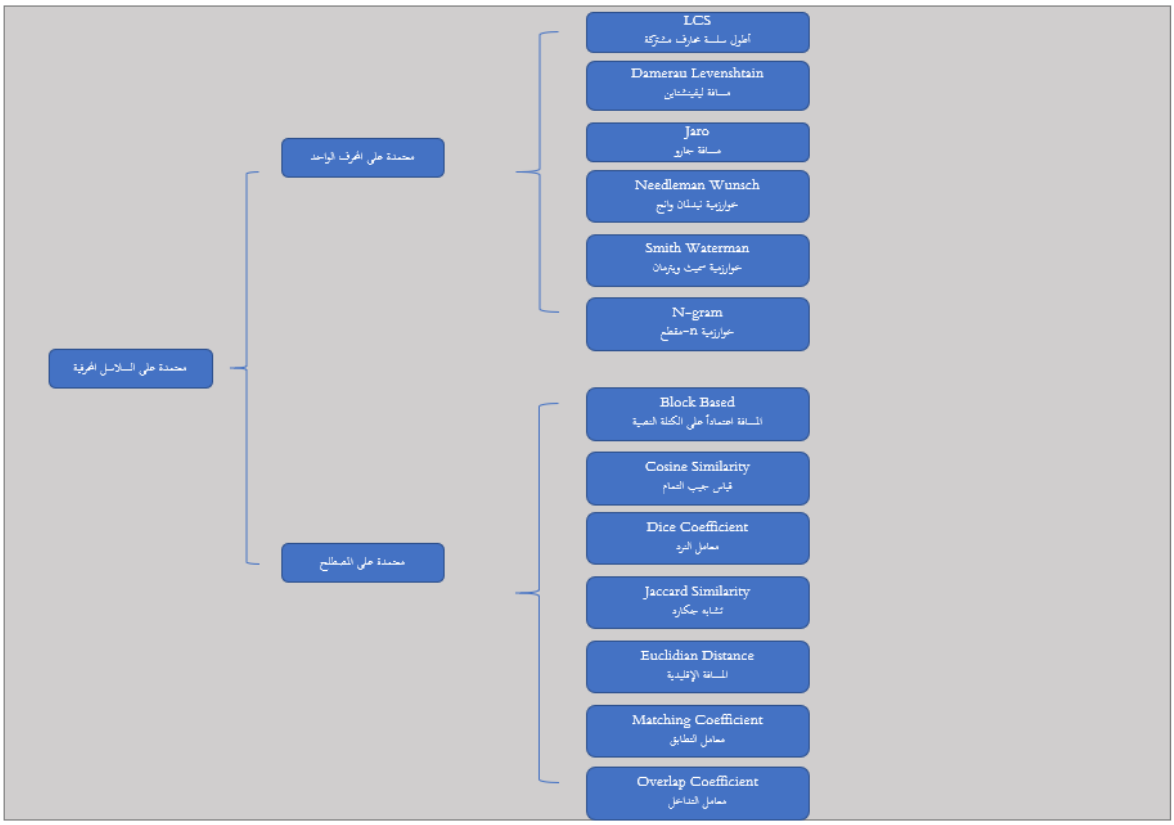
3.قياس تشابه النصوص

وتعني قياس مدى تشابه نصين وتقاربها معاً. تلعب مقاييس تشابه النصوص دوراً هاماً في العديد من الأبحاث والتطبيقات المتعلقة بمجال معالجة النصوص، مثل استخراج المعلومات Information Retrieval، وتصنيف النصوص Text Classification، والإجابة الآلية QAS، والترجمة الآلية Machine Translation، والتلخيص الآلي Text Summarization وغيرها الكثير. يقاس تشابه النصوص بتشابه الكلمات الواردة ضمنها، ويمكن أن يكون التشابه نحوياً أو دلالياً [9]. نستعرض فيما يلي كيفية قياس التشابه بين النصوص.

1.3.التشابه على المستوى النحوي

تعتمد مقاييس التشابه النحوي على شكل الكلمات في قياس التشابه بينهما، فيمكن مثلاً اعتبار كلمتين متشابهتين في حال كان لهما الجذع نفسه أو الجذر نفسه مثل كلمتي "سنلعب" و "لعِبوا"، هما متشابهتان لامتلاكهما الجذر نفسه "لعب". هناك عدة مقاييس للتشابه المعجمي بين الكلمات ويطلق عليها String based Similarity Metrics[1] ويوضح الشكل(3) بعض هذه الخوارزميات.

[1] التشابه القائم على السلسلة النصيّة.



الشكل 3 – بعض مقاييس التشابه على المستوى المعجمي[1] .

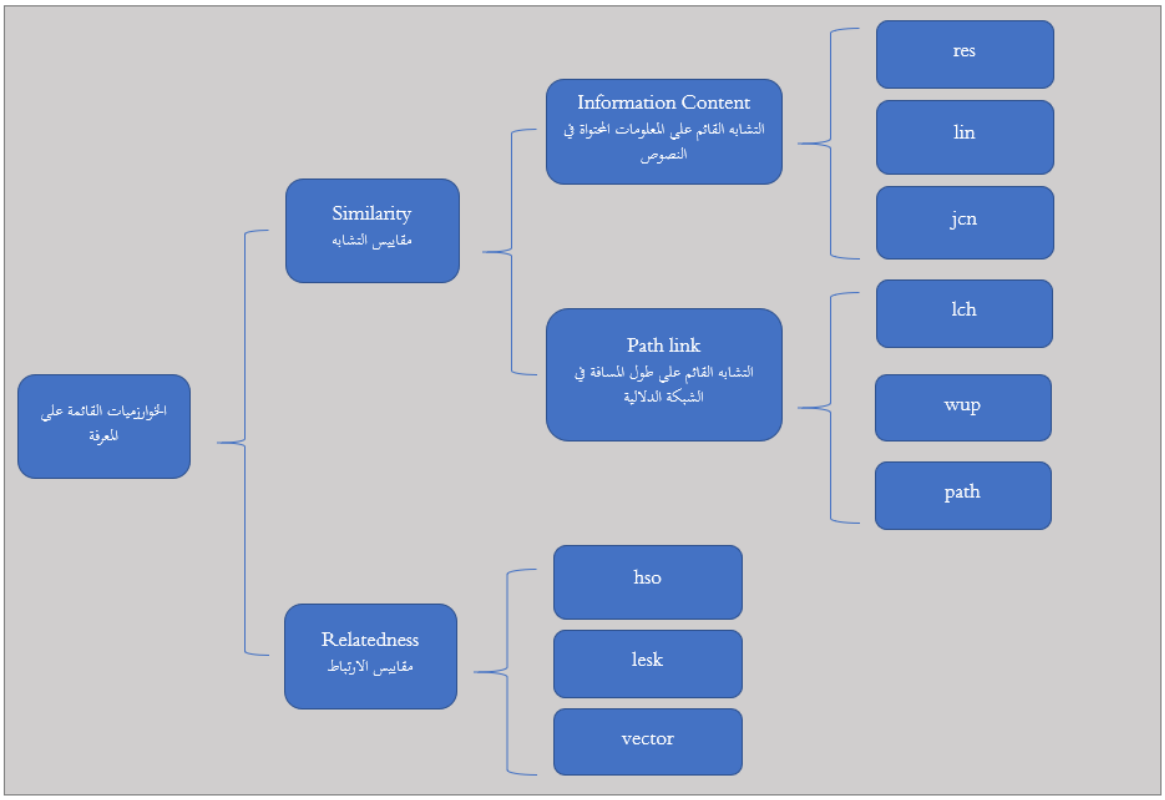
2.3.التشابه على المستوى الدلالي

يتم قياس التشابه بين الكلمات على المستوى الدلالي بأخذ معاني الكلمات بالاعتبار، بحيث تعتبر الكلمتين متشابهتين عندما تعبران عن معنى واحد، مثال كلمتي "مهنة" و "عمل". ويمكن قياس التشابه الدلالي بين الكلمات إما بالاعتماد على المعرفة او بالاعتماد على المدونات (بالاعتماد على السياق).

1.2.3.التشابه الدلالي القائم على المعرفة

التشابه القائم على المعرفة هو مقياس تشابه دلالي يحدد درجة التشابه بين الكلمات باستخدام المعلومات المستمدة من الشبكات الدلالية Wordnet. يبين الشكل(4) بعض الخوارزميات التي تقيس التشابه الدلالي القائم على المعرفة.

[1] لمعرفة المزيد عن هذه الخوارزميات يمكن الاطلاع على المرجع [9].

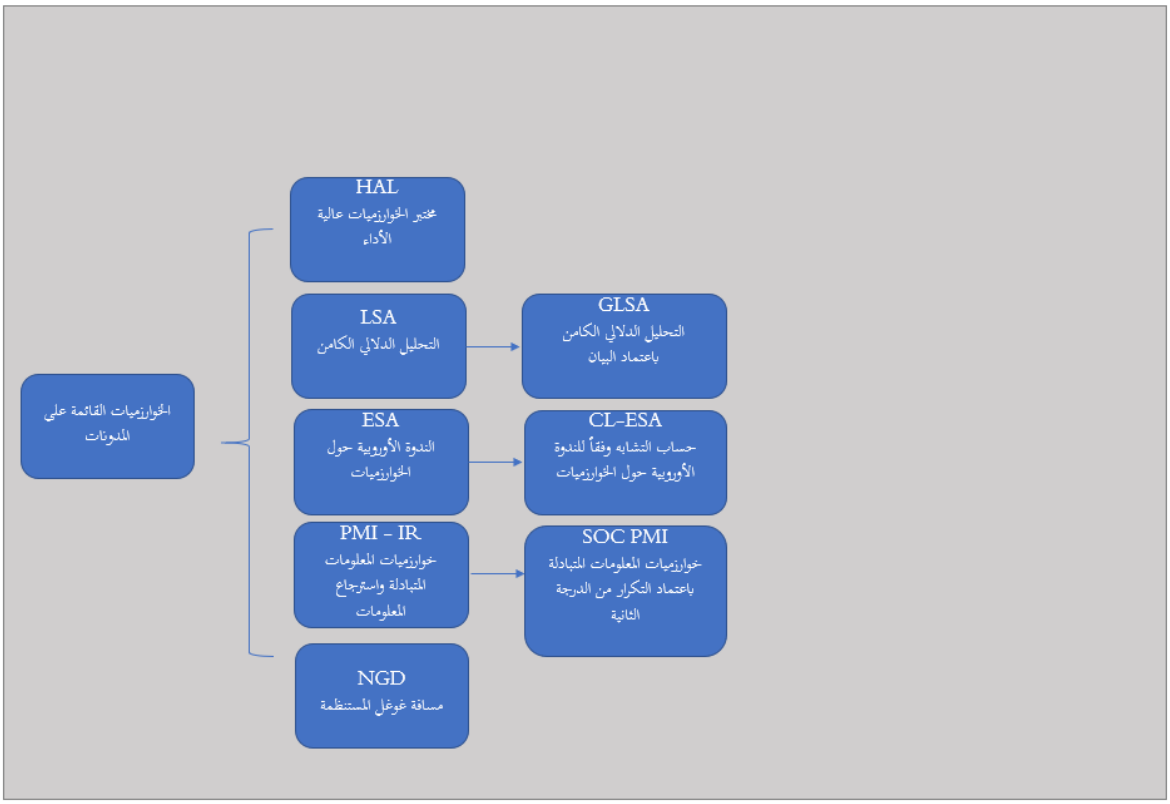


الشكل 4 – بعض مقاييس التشابه الدلالي القائم على المعرفة[1] .

2.2.3.التشابه الدلالي القائم على المدونات

التشابه القائم على المدونات هو مقياس تشابه دلالي يحدد التشابه بين الكلمات وفقًا للمعلومات المكتسبة من مجموعة كبيرة من المدونات النصية. يوضح الشكل(5) بعض مقاييس التشابه الدلالي القائم على المدونات.

[1] لمعرفة المزيد عن هذه الخوارزميات يمكن الإطلاع على المرجع [9].



الشكل 5 – بعض مقاييس التشابه الدلالي القائم على المدونات[1] .

قمنا باستخدام مفهوم تشابه النصوص في بحثنا في مكون استخراج المستندات ومكون استخراج المقاطع النصية. سنقوم بشرح التفاصيل في فصل النظام المقترح ضمن هذه الأطروحة.

4.تضمين الكلمات

يعد تضمين الكلمات تقنية حديثة في مجال معالجة اللغات الطبيعية، تعتمد على تمثيل الكلمة بمتجه رقمي، وذلك لتسهيل عمليات معالجة النصوص (مثال قياس التشابه بين النصوص يصبح عبارة عن قياس البعد بين المتجهات الرقمية الممثلة لها)، ويجري إنشاء هذه المتجهات اعتماداً على مدونات نصية كبيرة، بحيث تأخذ السياقات المختلفة للكلمة من المدونات النصية وتعطي المتجه الرقمي الممثل للكلمة. تمتلك الكلمات المتشابهة في سياقها تمثيلاً شعاعياً متقارباً أيضاً، كما في كلمتي "أستاذ" و "مدرس". تختلف التقنيات المستخدمة لإجراء عملية الربط بين الكلمة والمتجه الرقمي الممثل لها، فيمكن أن تتم باستخدام الشبكات العصبونية Neural Networks، أو تقنية تخفيض الأبعاد Dimensionality Reduction، أو باعتماد مصفوفة التواجد المشترك Co-Occurrence Matrix، أو باستخدام نماذج إحصائية وغيرها الكثير من الطرائق [10]. يمكننا استخدام

[1] لمعرفة المزيد عن هذه الخوارزميات يمكن الاطلاع على المرجع [9].

تضمين الكلمات في العديد من المهام المرتبطة بمعالجة اللغات الطبيعية مثل التحليل الصرفي، وتحليل المشاعر، وقياس تشابه النصوص، والإجابة الآلية وغيرها من التطبيقات.

هناك العديد من نماذج تضمين الكلمات، وتختلف باختلاف طريقة تدريبها وتمثيلها لمتجه الكلمة، فمنها النماذج المعتمدة على السياق Contextual Dependent Word Embeddings Models، ومنها النماذج المستقلة عن السياق Context Independent Word Embeddings Models، فيما يلي مقارنة بين هذين النوعين من النماذج.

1.4.نماذج تضمين الكلمات المستقلة عن السياق

تعتمد هذه النماذج على ربط كل كلمة من كلمات مدونة التدريب بمتجه رقمي وحيد، بحيث يكون دخل هذه النماذج هو الكلمة والخرج هو المتجه الرقمي الممثل لها، وبالتالي لا تنظر لسياق الكلمة أثناء التدريب، مثال كلمة "ذهب" فهي تمتلك تمثيلاً رقمياً وحيداً على الرغم من إمكانية ورودها بسياقات مختلفة كما في "ذهب الولد"، و"ارتفع سعر الذهب". يعتبر نموذج word2vec مثالاً عن هذه نماذج تضمين الكلمات المستقلة عن السياق [11].

- نموذج word2vec

وهي شبكة عصبونية مؤلفة من طبقتين تستخدم للتدريب على تضمين الكلمات، يتمثل دخلها بمدونات نصية كبيرة، ويتمثل الخرج بمتجهات رقمية تمثل كلمات هذه المدونات، ويعتبر هذا النوع من نماذج تضمين الكلمات المستقلة عن السياق. هناك نماذج word2vec مدربة مسبقاً ومتاحة للاستخدام في العديد من تطبيقات معالجة اللغات الطبيعية، ومنها ما يتوفر باللغة العربية مثل AraVec[1]، والتي تحتوي على ستة نماذج تضمين كلمات مختلفة، بحيث تم تجميع بيانات التدريب من ثلاثة مصادر أساسية وهي: تغريدات تويتر، وصفحات الوب المتنوعة، ومقالات الويكيبيديا. يبلغ عدد الكلمات التي يمكن تمثيلها باستخدام نماذج AraVec أكثر من 3,300,000,000 كلمة [12].

2.4.نماذج تضمين الكلمات المعتمدة على السياق

تختلف هذه النماذج عن النماذج المستقلة عن السياق، بحيث يمكن للكلمة الواحدة أن تتمثل بأكثر من متجه رقمي، وذلك وفقاً للسياق الواردة ضمنه، فيكون بذلك دخل هذه النماذج هو جملة كاملة والخرج هو المتجهات الرقمية لكلمات الجملة، وذلك لأنها تحتاج للسياق أثناء عملية التدريب، مثال كلمة "فأرة" فهي تمتلك تمثيلاً شعاعياً في جملة "فأرة الحاسوب" وآخر في

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

مثال "أكل القط الفأرة". يعتبر نموذج BERT(Bidirectional Encoder Representations from Transformers) مثالاً عن نماذج تضمين الكلمات المعتمدة على السياق [11].

1.2.4.نموذج BERT

BERT اختصاراً لـ Bidirectional Encoder Representations from Transformers، وهو نموذج لغوي يجري تدريبه باستخدام مدونات نصيّة كبيرة، وذلك بهدف تحقيق تضمين الكلمات لتستخدم لاحقاً في العديد من مهام معالجة اللغات الطبيعية، حيث تعتبر BERT من النماذج المعتمدة على السياق، وجدتها Google، واستخدمت في العديد من مهام معالجة اللغات الطبيعية وأثبتت كفاءتها في الفترة الأخيرة [13].

يجري تحقيق النموذج اللغوي BERT على مرحلتين وهما: مرحلة التدريب Bert Pretraining ،ومرحلة التخصيص Bert Fine Tuning، حيث تهتم مرحلة التدريب بفهم اللغة من حيث السياق (ماهي اللغة التي يتم التدريب عليها؟ وما هو سياقها؟)، في حين تهتم مرحلة التخصيص بإجراء مهمة محددة باستخدام النموذج اللغوي مثل تحقيق إجابة آلية، أوتصنيف المشاعر، أو التعرف على الكيانات المسماة وغيرها الكثير من المهام الممكن إنجازها.

مرحلة تدريب النموذج اللغوي Bert Pretraining

الهدف الأساسي من هذه المرحلة هو فهم اللغة التي يتم التعامل معها وذلك من خلال فهم سياقها، يجري ذلك بإنجاز عمليتي تدريب دون معلم unsupervised learning بآن واحد وهما: **نمذجة اللغة المقنعة Masked Language Modeling(MLM)**، و **التنبؤ بالجملة التالية Next Sentence Predication(NSP)**.

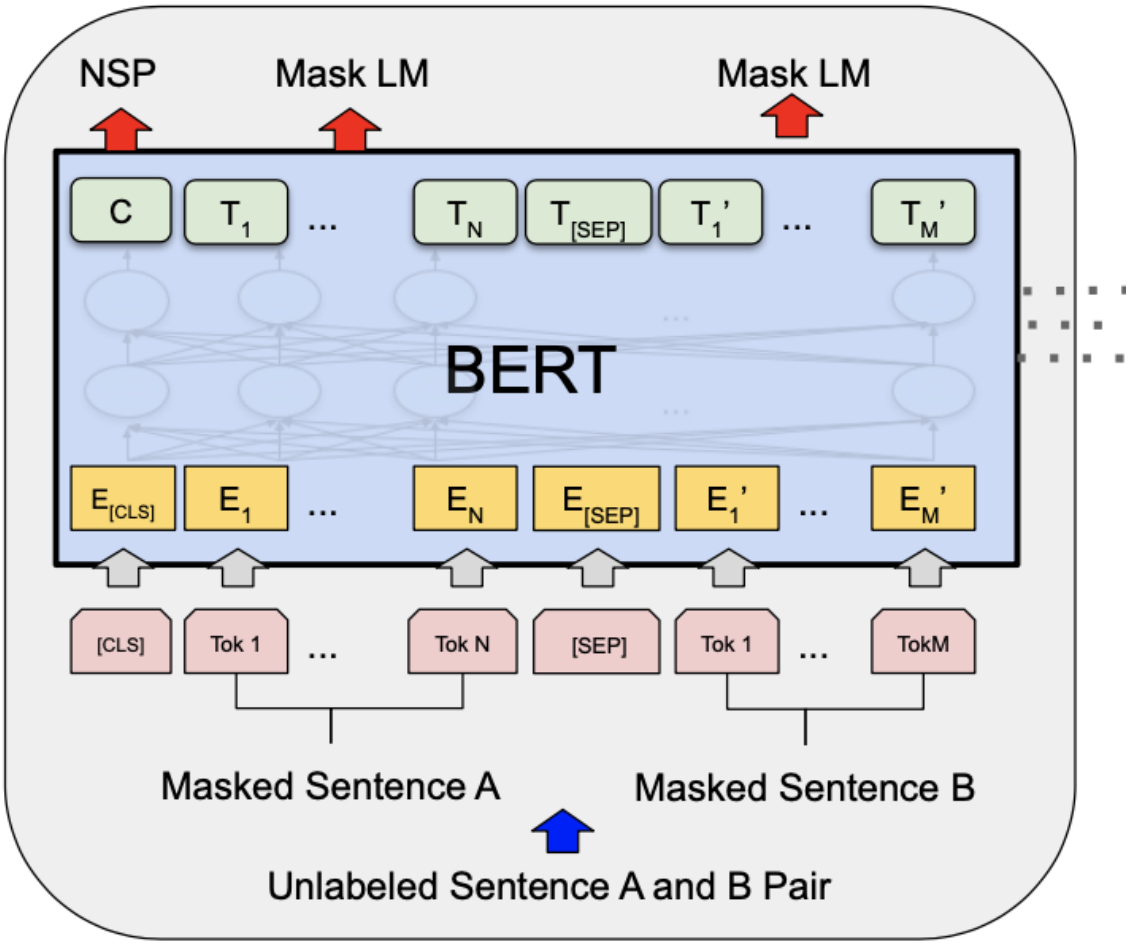
نمذجة اللغة المقنعة: يجري التدريب بإدخال جمل نصية على النموذج مع تقنيع بعض الكلمات من هذه الجمل، ليتم التنبؤ بالكلمات المقنعة وإعادتها بالخرج.

التنبؤ بالجملة التالية: ويتم بإدخال جملتين على النموذج (جملة أولى A، وجملة ثانية B)، ليتم التنبؤ ما إذا كانت الجملة B تتبع الجملة A في سياق اللغة أم لا (0 أو 1).

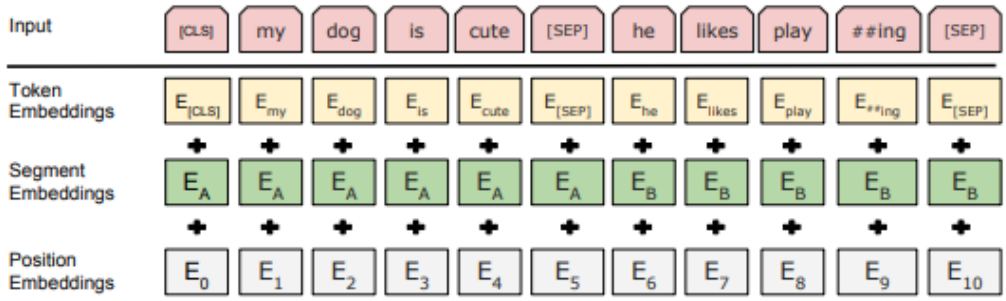
من خلال التدريب على المهمتين السابقتين يستطيع النموذج Bert فهم اللغة وبالتالي تضمين كلمات هذه اللغة تبعاً للسياق الذي وردت ضمنه.

قبل البدء بعمليات التدريب السابقة يجري تقطيع نص الدخل إلى مجموعة من الرموز tokens وفق طريقة Workpieces Tokenization، بحيث تجري عملية التقطيع وفقاً لرموز اللغة، مثال في اللغة العربية يجري تقطيع كلمة "قلنا" لجزئين وهما "قل" و "نا"، وتستند عملية التقطيع هذه على معجم يشمل كافة الرموز الممكنة للغة.

يوضح الشكل (6) عملية تدريب النموذج اللغوي، يتألف دخل النموذج من سلسلة نصية تمثل جملة واحدة في اللغة أو عدة جمل مفصولة برمز خاص [SEP]، يجري تقطيع سلسلة الدخل وفقاً للطريقة السابقة (Workpieces Tokenization) ويجري تمييز أول رمز token في هذه السلسلة برمز خاص [CLS] يتم التعامل معه لاحقاً في عملية التدريب، تتم لاحقاً ثلاثة عمليات تضمين متتالية على الدخل، انظر الشكل(7)، وهي: تضمين رموز الدخل Token Embeddings وذلك بالاستعانة بمعجم رموز اللغة، وتضمين الجمل Sentence Embeddings وذلك بهدف أخذ السياق بالاعتبار، وتضمين موقع الرمز Position Embeddings وذلك بهدف تمييز مكان الرمز في الجملة. تمر الأشعة الرقمية الناتجة عن مراحل التضمين الثلاثة عبر عدة محولات ترميز ثنائية الاتجاه موّزعة على عدة طبقات، بحيث يمثل خرج المحول دخلاً للمحول الذي يليه في الطبقة التالية يدعى المحول الواحد في كل طبقة كتلة ترميز Encoder Block. ويتمثل خرج النموذج بعدة رموز، انظر الشكل(6)، حيث يمثل الرمز NSP نتيجة مهمة التنبؤ بالجملة التالية؛ فيأخذ القيمة 1 في حال كانت الجملة B تتبع الجملة A في السياق وإلا يأخذ القيمة 0. إضافةً لمجموعة من الرموز $T_1, T_2, \ldots, T_n$ والتي تمثل نتيجة التنبؤ بالكلمات المقنعة[14].



الشكل 6 – البنية العامة لنموذج Bert في مرحلة التدريب.

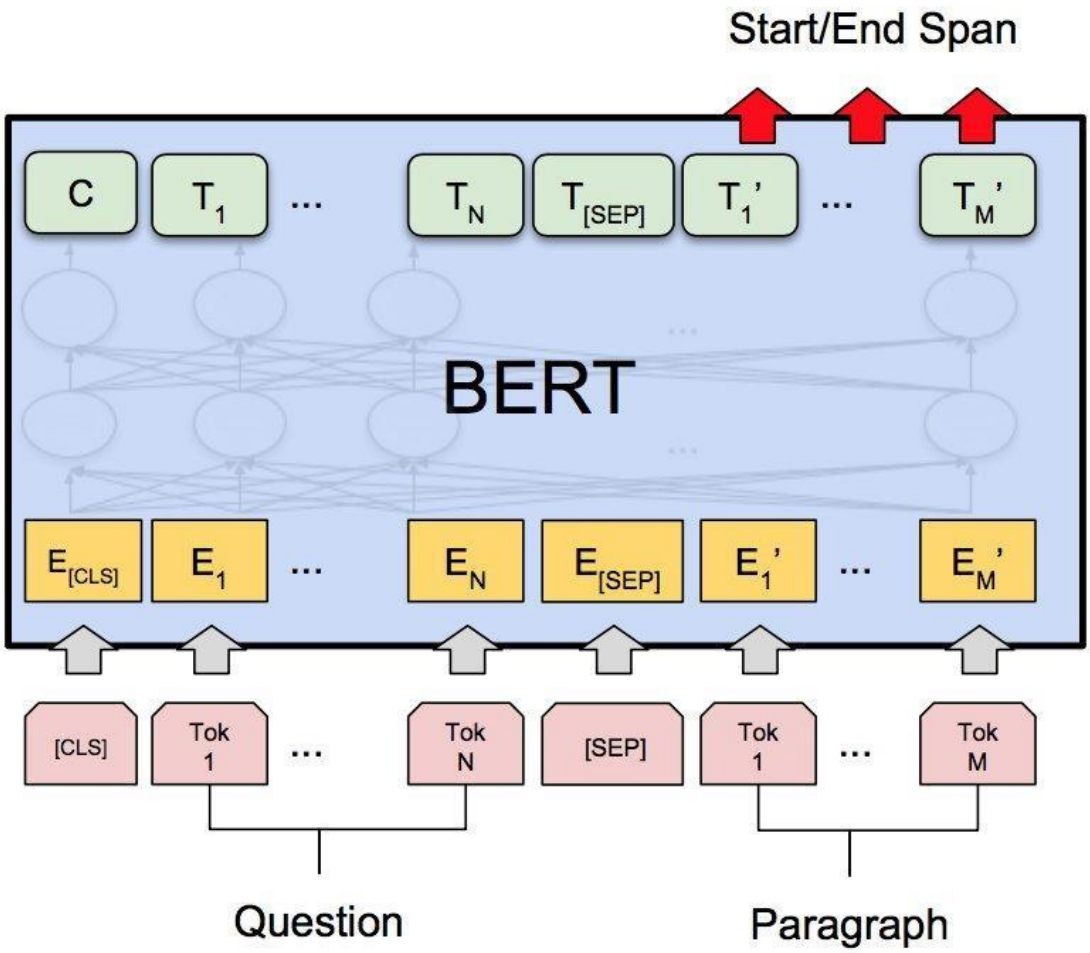


الشكل 7 – تمثيل دخل نموذج Bert.

مرحلة تخصيص النموذج اللغوي Bert Fine Tuning

يتم عادةً تخصيص نماذج BERT للقيام بمهمة محددة مثل تصنيف النصوص، وتحليل المشاعر وغيرها، وتسمى هذه العملية BERT Fine Tuning، وذلك بإضافة طبقات إضافية على خرج النموذج، يتم تدريب هذه الطبقات للقيام بالمهمة المحددة، بحيث تأخذ كدخل نتيجة تضمين الكلمات التي تعيدها BERT وتعطي بالخرج النتيجة الموافقة للمهمة المحددة، مثال في مصنفات تحليل المشاعر، تضاف طبقات المصنف على شبكة BERT بحيث تعطي تصنيفاً لجملة معينة على أنها إيجابية، أو سلبية، أو محايدة اعتماداً على تضمين كلمات الجملة الناتج عن خرج BERT. هناك العديد من المهام الأخرى التي يمكن تخصيص BERT بها، منها الإجابة الآلية، والتعرف على الكيانات المسماة [19].

يوضح الشكل(8) كيفية تخصيص نموذج Bert في مهمة الإجابة الآلية، حيث يتألف الدخل من نص السؤال والمقطع النصي الحاوي على الإجابة Paragraph، ويضاف لطبقة الخرج مجموعة من الرموز $T_1', T_2', \ldots, T_m'$ تمثل جزءً نصياً span مقتطعاً من المقطع النصي Paragraph كإجابة عن السؤال المطروح[13].



الشكل 8 - تخصيص نموذج Bert في الإجابة الآلية.

يتوفر العديد من نماذج BERT مدربة مسبقاً ويمكن استخدامها لإنجاز المهام المختلفة في معالجة اللغات الطبيعية، ومنها ما يتوفر باللغة العربية مثل AraBert[1] [15].

يوضح الجدول (1) مقارنة بين النماذج المستقلة عن السياق والنماذج المعتمدة على سياق وفقاً لعدة معايير وهي: السياق أثناء التدريب، وترتيب الكلمات، وطريقة التضمين، وتمثيل الكلمات غير الواردة ضمن مدونات التدريب.

[1] https://github.com/aub-mind/arabert

المعيار	النماذج المستقلة عن السياق	النماذج المعتمدة على السياق
السياق أثناء التدريب	لا تأخذ بالاعتبار سياق الكلمة أثناء التدريب، فدخلها كلمة وحيدة وخرجها متجهاً رقمياً يوافق الكلمة.	يختلف تمثيل الكلمة باختلاف السياق الواردة ضمنه، فدخل هذه النماذج هو الجملة كاملة والخرج هو التمثيل الشعاعي للكلمات المكونة لها.
ترتيب الكلمات	لا تهتم بترتيب الكلمة ضمن الجملة على اعتبار انها نماذج لا تنظر للسياق.	تقوم هذه النماذج بالاعتماد على معلومة موقع الكلمة ضمن الجملة index، أثناء قيامها بعملية إيجاد الشعاع الممثل للكلمة.
طريقة التضمين	تقابل كل كلمة بشعاع رقمي وحيد.	يتم تدريب هذه النماذج بحيث تعطي تمثيلاً مختلفاً للكلمة وفقاً للسياق الواردة ضمنه.
تمثيل الكلمات غير الواردة ضمن مدونات التدريب	في حال عدم ورود كلمة ما ضمن مدونات التدريب؛ فلا يمكن لهذه النماذج إيجاد تمثيلاً شعاعياً لها.	لاتنحصر هذه النماذج بتمثيل الكلمات الواردة ضمن مدونات التدريب فقط، وذلك لأنها تتدرب على مستوى أجزاء الكلمات workpieces.
مثال	Word2Vec	BERT

الجدول 1 – مقارنة بين النماذج المستقلة عن السياق والنماذج المعتمدة على السياق في عملية تضمين الكلمات [11].

قمنا باستخدام مفهوم تضمين الكلمات في عدة مراحل ضمن بحثنا، حيث قمنا باعتماد النماذج المستقلة عن السياق(AraVec) في مرحلتي توسيع السؤال استخراج المقاطع النصية، كما قمنا باعتماد النماذج المعتمدة على سياق(Bert) في مرحلة استخراج الإجابة. سنقوم بشرح التفاصيل في فصل النظام المقترح ضمن هذه الأطروحة.

5.الخاتمة

قمنا في هذا الفصل بعرض دراسة نظرية مبسّطة عن بعض المفاهيم التي يعتمد عليها بحثنا بشكل أساسي، بحيث تكلّمنا في البداية عن مجال معالجة اللغات الطبيعية ومستويات المعالجة التي تمر بها، وتطرقنا لشرح ضرورة تطوير معالجة اللغة العربية. استعرضنا أيضاً شرحاً لمفهومي تشابه النصوص و تضمين الكلمات وأهميته في مجال معالجة اللغات الطبيعية واستخدامه في نظامنا بشكل خاص، وسنبين في الفصل الثالث دراسة للنظم المشابهة ومقارنتها فيما بينها، إضافةً لتحديد النقاط التي سيرتكز عليها البحث انطلاقاً من ثغرات هذه النظم.

الفصل الثالث: دراسة الأعمال ذات الصلة في نظم الإجابة الآلية

1.معايير المقارنة

قبل البدء بعرض أهم الأعمال في مجال الإجابة الآلية، لابد في البداية من تحديد بعض المعايير لمقارنة هذه الأعمال وفقها. من هذه المعايير: اللغة المعالجة، ومجال المعطيات، وأنماط الأسئلة المعالجة (الدخل)، وشكل الخرج، وآلية الاختبار والنتائج وتكلفة التنفيذ.

1.1.اللغة المعالجة

تختلف نظم الإجابة الآلية باختلاف اللغة التي تعالجها، فهناك الكثير من النظم التي تهتم بمعالجة اللغات الأجنبية مثل اللغة الإنكليزية، وهي الأكثر كفاءة مقارنة بتلك التي تهتم بمعالجة اللغة العربية، ويعود ذلك لعدة أسباب منها انتشار اللغة الإنكليزية وسهولة معالجتها مقارنة بالعربية.

2.1.مجال المعطيات

يمكن لنظم الإجابة الآلية أن تكون محددة المجال أو مفتوحة المجال، تبعاً لمجال المعطيات التي تتعامل معها. في النظم المحددة المجال، يستطيع النظام الإجابة عن أسئلة ضمن مجال محدد مثل نظم الاستفسارات الطبية أو نظم الإجابة الآلية الخاصة بالقرآن الكريم وغيرها، بينما لا تنحصر النظم المفتوحة المجال بمجال إجابة محدد، إذ يمكنها تقديم إجابة عن كافة الاستفسارات على اختلاف مجالاتها؛ فيمكن مثلاً للنظام الواحد أن يجيب عن أسئلة ضمن الرياضة، والصحة، والفن، والاقتصاد وغيرها.

3.1.أنماط الأسئلة المعالجة (الدخل)

تنحصر بعض النظم في الإجابة عن أسئلة من نمط محدد، فهناك نظم تجيب عن أسئلة من نوع "لماذا" فقط، وهناك نظم أيضاً تهتم بالأسئلة الزمنية مثل "متى" أو "في أي عام"، يوضح الجدول (2) بعض الأنماط الرئيسية للأسئلة الممكن معالجتها. يمكن للنظام الإجابة عن أنماط مختلفة من الأسئلة دون أن يقتصر على نمط وحيد معين.

نمط السؤال	شرح	مثال
الأسئلة التعريفية **Factoid Questions**	وهي الأسئلة التي تستخدم للاستفسار عن شخص أو مكان أو زمان أو منظمة وغيرها، ويتوقع الإجابة عنها باسم الشخص أو باسم مكان أو بتاريخ.	"أين عاش أحمد عمر خاشقجي؟"
أسئلة التعداد **List Questions**	وهي الأسئلة التي يجاب عنها بقائمة من الحقائق أو الأسماء، ويمكن القول أنها مثل الأسئلة التعريفية ولكن يجاب عنها بقائمة من الأجوبة عوضاً عن الجواب الوحيد.	"ما هي المهن التي عمل بها خاشقجي؟"
أسئلة التأكيد **Confirmation Questions**	وهي الأسئلة التي يجاب عنها بنعم أو لا.	"هل قرة القدم رياضة شعبية؟"
أسئلة التلخيص **Summarization Questions**	وهي الأسئلة التي يجاب عنها بملخص أو شرح بسيط مثل أسئلة الاستفسار "كيف" و "لماذا" وغيرها.	"لماذا سميت غزوة بدر بهذا الاسم؟"
أسئلة الفرضيات **Hypothetical Questions**	وهي الأسئلة التي يستفسر بها عن الشيء في حال افتراض وقوع حدث معين.	"إذا امطرت غداً ماذا سيحدث للزرع؟"

الجدول 2 –بعض الأنماط الرئيسية للأسئلة.

4.1.شكل الخرج

تختلف نظم الإجابة الآلية في الشكل النهائي للخرج، فهناك نظم تكتفي باستخلاص المقطع النصي الحاوي على الإجابة والذي يمكن أن يكون جملة واحدة أو عدة جمل وتعيده كإجابة نهائية، بينما تعيد نظم أخرى إجابةً محددة وواضحة للسؤال المطروح، وذلك اعتماداً على نمط السؤال. مثال: إذا كان السؤال المطروح من نمط أسئلة التأكيد فيعيد النظام إجابة محددة وفقاً لهذا النمط وهي نعم أو لا.

5.1.آلية الاختبار والنتائج

هناك عدة معايير لقياس جودة نظم الإجابة الآلية منها Precision(P)، وRecall(R)، وF-Measure(F1)، و Exact Match(EM)، و Sentence Match(SM) والدقة Accuracy. ويجري حساب القياسات السابقة وفقاً للمعادلات التالية [5,3]:

$$P = \frac{\text{عدد الإجابات الصحيحة} \cap \text{عدد الإجابات المستخرجة}}{\text{عدد الإجابات المستخرجة}}$$

$$R = \frac{\text{عدد الإجابات الصحيحة} \cap \text{عدد الإجابات المستخرجة}}{\text{عدد الإجابات الصحيحة الممكنة في مجموعة البيانات}}$$

$$F1 = \frac{\text{عدد الكلمات المتقاطعة مع الإجابة الصحيحة}}{\text{عدد الكلمات الكلي}}$$

$$SM = \frac{\text{عدد الإجابات المتوافقة مع الإجابة الصحيحة في النص الواردة ضمنه}}{\text{عدد الإجابات الكلي}}$$

$$EM = \frac{\text{عدد الإجابات المتوافقة تماما مع الإجابة الصحيحة}}{\text{عدد الإجابات الكلي}}$$

$$Acc = \frac{\text{عدد الإجابات الصحيحة (بعد مراجعة بشرية)}}{\text{عدد الإجابات الكلي}}$$

6.1. التكلفة

تقاس تكلفة نظم الإجابة الآلية بمدى حاجة المنهجية المقترحة لعتاد قوي (GPU/TPU)، إضافة لزمن التدريب في حال كانت الطريقة المقترحة تعتمد على تقنيات التعلم العميق وتحتاج مرحلة تدريب.

2.أهم الأعمال المشابهة

1.2.الأعمال المشابهة باللغات الأجنبية

1.1.2. DrQA – 2017

وهي اختصار لـ Document Retriever | Document Reader Question Answering

- **لغة المعالجة**

يهتم النظام بالإجابة عن الأسئلة المطروحة باللغة الإنكليزية.

- **مجال المعطيات**

ليس هناك مجال محدد ينحصر النظام به فهو نظام إجابة آلية مفتوح المجال ويعتمد في استخراج الإجابة على مقالات الويكيبديا المختلفة.

- **أنماط الأسئلة المعالجة**

يهتم النظام بالإجابة عن الأسئلة التعريفية Factoid questions.

- **آلية العمل المتّبعة**

يتألف النظام بشكل عام من مكوّنين أساسيين [16] (انظر الشكل(9)):

- مكون استخراج المستندات Document Retriever:

يجري في هذا المكون استخراج المستندات (مقالات الويكيبيديا) المرتبطة بالسؤال المطروح، جرى ذلك عن طريق تمثيل كل من المستند والسؤال المطروح بمتجه موزون وفق قياس TF-IDF، ومن ثم حساب قيمة التشابه بين هذه المتجهات باستخدام مقياس جيب التمام، واعتماد أكثر خمسة مستندات تشابهاً مع نص السؤال كدخل للمرحلة اللاحقة، ولتحسين نتائج هذه المرحلة جرى أخذ ترتيب الكلمات بالاعتبار وذلك عن طريق أخذ مجموعة من الكلمات n-grams كسمات للمتجه الممثل للمستندات بدلاً من اعتماد كلمة وحيدة، وكذلك الأمر بالنسبة لمتجه السؤال.

- مكون استخراج الإجابة Document Reader:

يقوم النظام باقتطاع جزء نصي من نصوص المستندات ويعيده كإجابة نهائية عن السؤال المطروح، وجرى ذلك بتدريب شبكة عصبونية Recurrent Neural Network (RNN) دخلها متجهين رقميين q ويمثل متجه السؤال، و p ويمثل متجه مجموعة المقاطع النصية الناتجة عن المرحلة السابقة (نصوص المستندات)، وخرج الشبكة هو i و j حيث i هو الكلمة التي تبدأ بها الإجابة وj الكلمة التي تنتهي بها الإجابة المقتطعة من النصوص.

ترميز متجه السؤال

يتمثل السؤال المؤلف من l كلمة بمتجه رقمي $q = \{q_1, q_2, q_3, \ldots, q_l\}$، حيث يتم حساب q استناداً للمتجهات الناتجة عن تضمين الكلمات المكونة للسؤال (باستخدام نموذج تضمين كلمات مدرب مسبقاً)، وتوضح المعادلة التالية آلية حساب متجه السؤال q:

$$q = \sum_j bj\ qj$$

حيث b_j معامل يعبر عن أهمية الكلمة j في السؤال، و q_j المتجه الناتج عن تضمين الكلمات للكلمة j. ويتم الحصول على قيم معاملات الأهمية لكلمات السؤال نتيجة تدريب شبكة عصبونية منفصلة لهذا الغرض.

ترميز متجه مجموعة المستندات:

تتمثل مجموعة النصوص المعتمدة لاستخراج الإجابة بمتجه $p = \{p_1, p_2, p_3, \ldots, p_m\}$، بحيث تمثل p_i المتجه الرقمي للكلمة i في مجموعة النصوص. يجري في البداية إيجاد متجه السمات Features Vector من أجل كل كلمة من كلمات المستندات، ويجري حساب متجه الكلمة وفقاً لعدة سمات وهي: متجه تضمين الكلمات الممثل للكلمة، قيمة ثنائية (0 أو 1) تعبر عن كون الكلمة موجودة كما هي Exact Match في نص السؤال، قيمة ثنائية تعبر عن كون الكلمة موجودة كجذع Stem في نص السؤال، قيمة ثنائية تعبر عن كون الكلمة موجودة بعد التحويل لصيغة lower case في نص السؤال، Part Of Speech (POS) للكلمة، Named Entity Recognition (NER) للكلمة، Term Frequency(TF) للكلمة.

يجري استخراج متجه مجموعة النصوص p باعتماد متجهات كلماته وذلك بتدريب شبكة عصبونية منفصلة لهذا الغرض، دخلها متجهات كلمات النص وخرجها متجه رقمي يمثل النص ككل كما توضح المعادلة التالية:

$$\{\mathbf{p}_1, \ldots, \mathbf{p}_m\} = \mathrm{RNN}(\{\tilde{\mathbf{p}}_1, \ldots, \tilde{\mathbf{p}}_m\})$$

تدريب الشبكة العصبونية واستخراج الإجابة:

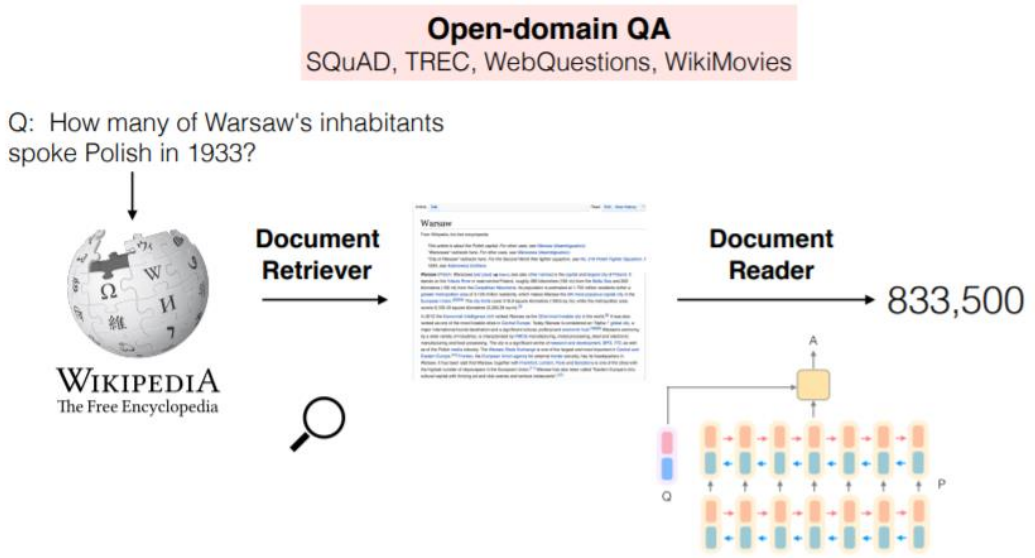
يجري تدريب الشبكة بحيث تعطي بالخرج كلمة البداية i وكلمة النهاية j للجزء النصي المقتطع كإجابة، بحيث يجري أثناء التدريب حساب احتمالية كل كلمة من كلمات مجموعة النصوص كونها كلمة بداية واحتمال كونها كلمة نهاية وفقاً للمعادلات التالية:

$$\begin{aligned} P_{start}(i) &\propto \exp(\mathbf{p}_i \mathbf{W}_s \mathbf{q}) \\ P_{end}(i) &\propto \exp(\mathbf{p}_i \mathbf{W}_e \mathbf{q}) \end{aligned}$$

حيث p_i متجه السمات للكلمة i و q متجه السؤال و W مصفوفة الأوزان للشبكة.

ويتم اختيار i و j بحيث تعطي أكبر قيمة ممكنة للعبارة التالية:

$$P_{start}(i) \times P_{end}(j)$$



الشكل 9 – الشكل العام لنظام DrQA.

- **شكل الخرج**

يعيد النظام جزءً نصيّاً مقتطعاً من مجموعة المستندات كإجابة نهائية عن السؤال المطروح.

- **آلية الاختبار والنتائج**

جرى تدريب النظام على مجموعة البيانات (SQuAD[1]) The Stanford Question Answering Dataset

واختباره على مجموعات بيانات مكونة من ثلاثة مجموعات جزئية مختلفة وهي:

(CuratedTREC, WebQuestions and WikiMovies)

وجرى تقييم النظام وفق معياري F1 و Exact Match(EM)، واستطاع النظام تحقيق EM = 70% و F1 = 79%.

- **التكلفة**

تعتبر المنهجية المقترحة في هذا النظام مكلفة نظراً لحاجتها لعتاد قوي ولزمن للقيام بعملية تدريب الشبكات العصبونية الثلاث (شبكة ترميز شعاع السؤال، شبكة ترميز شعاع مجموعة النصوص، شبكة الحصول على الإجابة).

[1] https://rajpurkar.github.io/SQuAD-explorer/

2.1.2. SimBioNLQA – 2020

- **لغة المعالجة**

يهتم النظام بالإجابة عن الأسئلة المطروحة باللغة الإنكليزية.

- **مجال المعطيات**

يجيب النظام عن الأسئلة المطروحة في مجال الطب الحيوي.

- **أنماط الأسئلة المعالجة**

يهتم النظام بالإجابة عن الأسئلة من الأنماط التالية:

- أسئلة التأكيد (نعم – لا).
- الأسئلة التعريفية.
- أسئلة التعداد.
- أسئلة التلخيص.

- **آلية العمل المتّبعة**

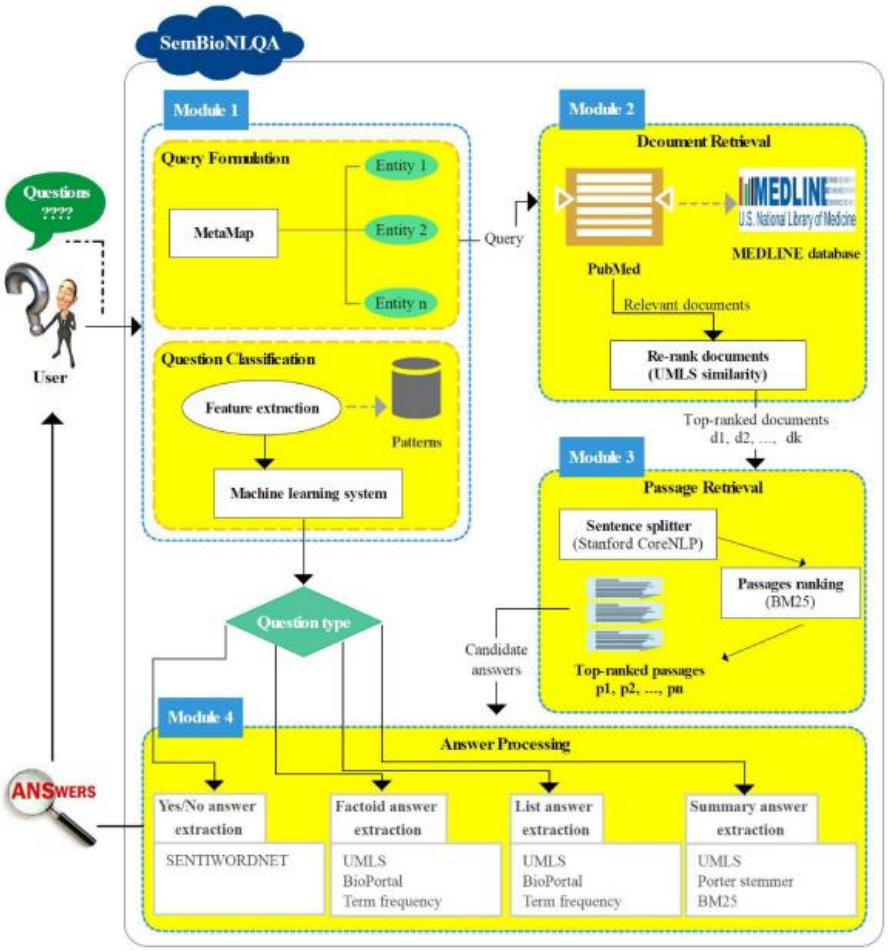
يتألف النظام من أربعة مكوّنات أساسية [7] وهي على التتالي مكوّن معالجة السؤال ومكوّن استخراج المستندات ومكوّن استخراج المقاطع النصيّة ومكوّن استخراج الإجابة، (انظر الشكل (10)). يتم ضمن مكوّن معالجة السؤال تصنيفه أولاً لأحد الأصناف الأربعة التي يعتمدها النظام ومن ثم إجراء مراحل المعالجة اللغوية النحوية على كلمات السؤال (التجذيع واستخراج نمط الكلام). يجري لاحقاً، في مكوّن استخراج المستندات، استخراج أكثر 5 مستندات ارتباطاً بسؤال المستخدم وذلك من خلال استخدام محرك بحث PubMed[1]. ثم اعتماداً على المستندات المستخرجة يتم استخراج المقاطع النصيّة ضمن مكون استخراج المقاطع النصيّة، ويتم ذلك بقياس التشابه بين السؤال والمقطع النصي، بعد إجراء معالجة لغوية نحوية على كل منهما، وباستخدام مقياس BM25[2]. تستخدم هذه المقاطع النصيّة لاستخراج الإجابة المحددة للسؤال المطروح وذلك ضمن مكوّن استخراج الإجابة الذي يعتبر المكوّن الأساسي في هذا النظام. يجري استخراج الإجابة تبعاً لنوع السؤال المطروح، وفي حال صُنف السؤال سؤال تأكيد يتم استخراج الإجابة من المقاطع النصيّة اعتماداً على معالجتها دلالياً لاستخراج المعنى المجمل لها (جملة سلبية أم إيجابية)، ووفقاً لذلك يجيب النظام بنعم أو لا. في حالة أسئلة التعريف والتعداد يتم معالجة المقاطع النصية باستخدام آلية التعرف على الكيانات المسمّاة Named Entity Recognition ويعيد النظام هذه الكيانات كإجابات نهائية وفقاً لنوع الكيان الذي تم

---

[1] https://pubmed.ncbi.nlm.nih.gov

[2] **BM25** (**Okapi BM25**): اختصاراً لكلمة أفضل مطابقة Best Matching، هو تابع ترتيب تستخدمه محركات البحث لتقدير مدى صلة المستندات باستعلام بحث معين.

السؤال عنه (مكان، شخص، تاريخ وغيره). أما بالنسبة لأسئلة التلخيص؛ فيكتفي النظام برد المقاطع النصيّة الناتجة عن المكوّن السابق كإجابة لهذا النوع من الأسئلة لكونها تشكل تلخيصاً لما تم السؤال عنه. وبذلك تجري المعالجة في النظام على كلا المستويين النحوي والدلالي: المستوى النحوي بهدف استخراج المستندات والمقاطع النصيّة المرتبطة بالسؤال وكذلك لاستخراج الإجابة، والمستوى الدلالي لتحديد الإجابة المحددة عن بعض أنماط الأسئلة.



الشكل 10 – البنية العامة لنظام SemBioNLQA.

- **شكل الخرج**

يعيد النظام جواباً محدداً وفقاً لنوع السؤال المطروح.

- **آلية الاختبار والنتائج**

تم اختبار النظام على مجموعة بيانات متاحة من قبل تحدي BioASQ[1] وذلك من أجل مقارنته بالنظم السابقة له في نفس التحدي لسنوات 2015 و 2016 و 2017، تم قياس دقة النظام وفق عدة مقاييس تختلف باختلاف نمط السؤال وهي Accuracy, Mean Reciprocal Rank (MRR), Precision, Recall, F-

[1] http://bioasq.org

Measure, Rouge[1]. تم تقسيم بيانات الاختبار لخمسة أقسام وقياس الدقة وفق المقاييس السابقة، ويبين الجدول (3) نتائج البحث من أجل أحد مجموعات البيانات والمعتمدة في تحدي عام 2015.

مجموعات البيانات	نوع السؤال					
	أسئلة محددة الإجابة					أسئلة غير محددة الإجابة
	أسئلة التأكيد	أسئلة تعريفية	أسئلة التعداد			أسئلة التلخيص
	Accuracy	MRR	Recall	F-Measure	Precision	Rouge - 2
مجموعة1	0.6970	0.1692	0.2409	0.1830	0.1545	0.2716
مجموعة2	0.6250	0.1776	0.2714	0.2172	0.1929	0.3123
مجموعة3	0.8612	0.1840	0.2927	0.2524	0.2353	0.3879
مجموعة4	0.7600	0.2690	0.2713	0.2588	0.2783	0.3917
مجموعة5	0.6071	0.1568	0.0736	0.0625	0.0583	0.3440

الجدول 3 - نتائج بحث SimBioNLQA

- **التكلفة**

تعتبر المنهجية المقترحة في هذا البحث غير مكلفة من ناحية العتاد والوقت؛ فليس هناك عمليات تدريب ضخمة تحتاج لمعالجات قوية، ويمكن تحقيق الخوارزميات المقترحة في البحث بزمن مقبول.

3.1.2. Efficient QA – 2021

- **لغة المعالجة**

يهتم النظام بالإجابة عن الأسئلة المطروحة باللغة الإنكليزية والفرنسية.

- **مجال المعطيات**

ليس هناك مجال محدد ينحصر النظام به فهو نظام إجابة آلية مفتوح المجال.

- **أنماط الأسئلة المعالجة**

لايهتم النظام بأنماط محددة من الأسئلة، فهو يجيب بقطع النظر عن نمط السؤال (الإجابة غير مرتبطة بنمط السؤال).

- **آلية العمل المتّبعة**

يجري في البداية إيجاد مجموعة من مقترحات الإجابات $A = \{A_1, A_2, \ldots, An\}$ وذلك من ضمن مجموعة من النصوص التي تشكل مجموعة بيانات النظام $C = \{C_1, C_2, \ldots, C_n\}$، بحيث يجري إيجاد الإجابات بقطع

[1] **Rouge**: هو مقياس يستخدم لقياس جودة نظم التلخيص الآلي، يعتمد في قياس الجودة على مقارنة الملخص الناتج عن النظام بملخصات مرجعية تم وضعها من قبل خبراء.

النظر عن السؤال المطروح؛ أي تجري عملية استخراج الإجابات كعملية معالجة سابقة لتشغيل النظام Off Line Processing، وعند طرح سؤال معين Q تتم مطابقته مع الإجابات المقترحة والمخزنة مسبقاً A ويعيد النظام أكثر إجابة متوافقة مع السؤال المطروح كما توضح المعادلة التالية:

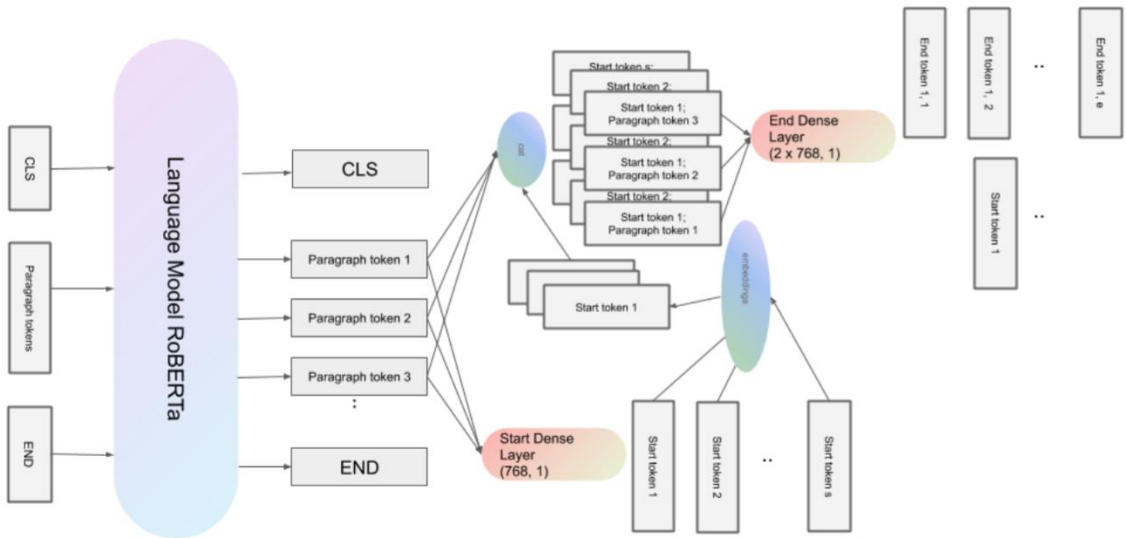
$$answer = argmax_{A_i}\, G(Q) \cdot H(A_i)$$

حيث answer الإجابة النهائية عن السؤال Q، و G(Q) هو السؤال بعد ترميزه باستخدام نموذج ترميز G، و $H(A_i)$ تمثل الإجابة المقترحة من مجموعة الإجابات بعد ترميزها باستخدام نموذج ترميز H [17].

توليد مجموعة الإجابات المقترحة A:

تتمثل الإجابة في هذا النظام بجزء نصي مقتطع من مجموعة النصوص المعتمدة، بحيث يتم تمثيل المقطع النصي (الإجابة) بكلمة بداية وكلمة نهاية. يجري إيجاد مجموعة الإجابات المقترحة استناداً إلى مجموعة من النصوص (مجموعة البيانات)، وتتم هذه العملية مرة واحدة فقط قبل تشغيل النظام. جرى تحقيق هذه المرحلة باستخدام نموذج تضمين كلمات معتمد على السياق BERT ، حيث جرى اعتماد نموذج مدرب مسبقاً وهو RoBERTa[1] يأخذ بالدخل كلمات النصوص tokens ويعيد نتيجة تضمين هذه الكلمات (متجهات رقمية)، وجرى تخصيص هذا النموذج بإضافة طبقتين إضافيتين للخرج، انظر الشكل(11)، تأخذ الطبقة الأولى نتيجة تضمين الكلمات كدخل لها، وتعيد مجموعة من الكلمات المقترحة s ككلمات بداية للإجابة، وتدخل مجموعة الكلمات هذه إضافة للدخل السابق إلى الطبقة الثانية التي تعيد مجموعة من الكلمات المقترحة e ككلمات نهاية، وبذلك يعيد النموذج المقترح s × e إجابة مقترحة.

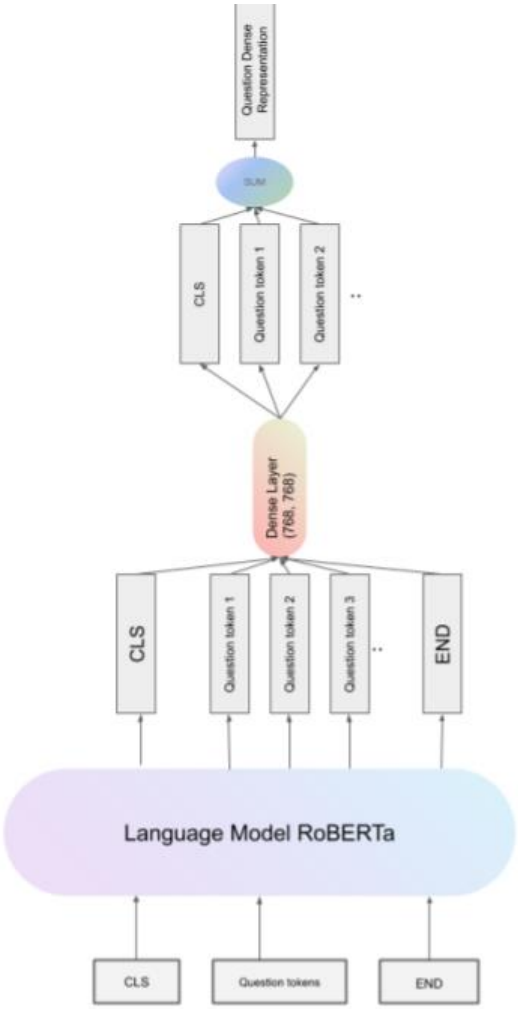
[1] https://github.com/pytorch/fairseq/tree/master/examples/roberta



الشكل 11 – منهجية توليد مجموعة الإجابات المقترحة في نظام EfficentQA.

ترميز نص السؤال G:

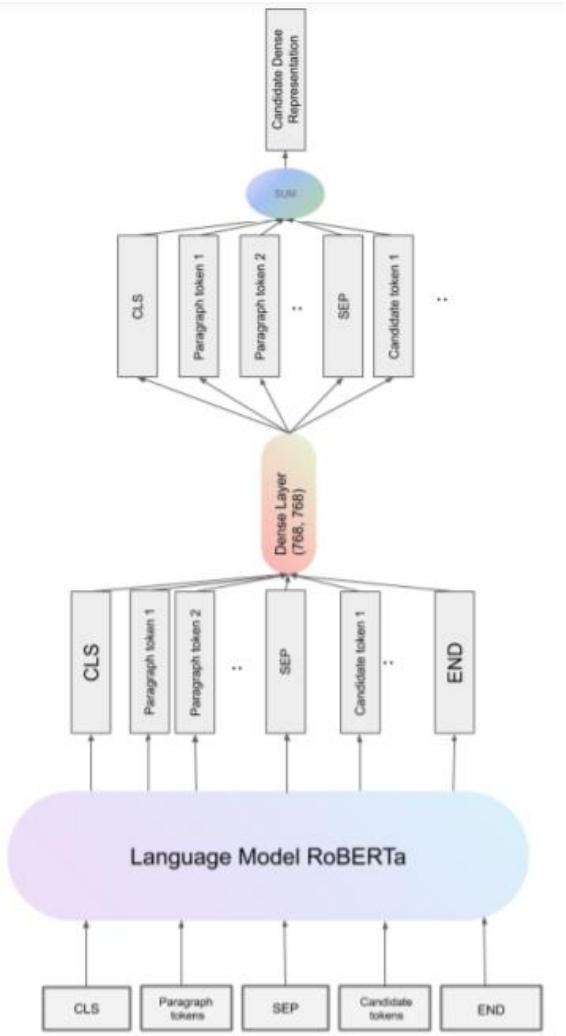
يتم ترميز السؤال باستخدام نموذج تضمين الكلمات المستخدم نفسه في مرحلة توليد مجموعة الإجابات المقترحة (RoBERTa)، حيث يتم تضمين كلمات نص السؤال ثم أخذ متوسط النتائج كنتيجة نهائية، كما يوضح الشكل(12).



الشكل 12- منهجية ترميز نص السؤال في نظام EfficentQA.

ترميز الإجابات H

يتم ترميز الإجابات باستخدام نموذج تضمين الكلمات المستخدم نفسه في مرحلة توليد مجموعة الإجابات المقترحة (RoBERTa)، حيث يتم تضمين كلمات الإجابة ثم أخذ متوسط النتائج كنتيجة نهائية، كما يوضح الشكل(13).



الشكل 13 – منهجية ترميز الإجابة في نظام EfficentQA.

- **شكل الخرج**

جزء نصي مقتطع من مجموعة النصوص التي تشكل مجموعة بيانات النظام.

- **آلية الاختبار والنتائج**

جرى اختبار النظام على مجموعة بيانات منمّطة باللغة الإنكليزية وهي SQuAD وكذلك مجموعة بيانات منمّطة باللغة الفرنسية وهي FQuAD1[1]، وتم تقييم النظام وفق معياريي Exact Match(EM) و F1 واستطاع النظام تحقيق النتائج EM = 92.3% و F1 = 96.7%.

[1] https://fquad.illuin.tech/

- **التكلفة**

هناك ثلاثة نماذج شبكية تحتاج لتدريب في المنهجية المقترحة وهي: نموذج إيجاد مجموعة الإجابات المقترحة، ونموذج ترميز نص السؤال G، ونموذج ترميز الإجابة H؛ وبالتالي هناك تكلفة من ناحية العتاد الذي تحتاجه عمليات التدريب، حيث تم التدريب باستخدام معالج 24GB GPU NVIDIA Quadro RTX 6000 ، إضافة لزمن التدريب الذي استغرق حوالي الأسبوع كما هو محدد في البحث.

2.2.الأعمال المشابهة باللغة العربية

1.2.2. EWAQ – 2015

اختصاراً لـ Entailment based Why Arabic Questions Answering.

- **لغة المعالجة**

يهتم النظام بالإجابة عن الأسئلة المطروحة باللغة العربية الفصحى.

- **مجال المعطيات**

لاينحصر النظام بمجال معطيات محدد فهو مفتوح المجال.

- **أنماط الأسئلة المعالجة**

يجيب النظام عن الأسئلة المطروحة من نمط "لماذا" فقط.

- **آلية العمل المتّبعة**

تمر عملية المعالجة لاستخراج الجواب المناسب بثلاث مراحل أساسية [18] وهي : مرحلة معالجة السؤال Question Processing، ومرحلة استخراج المقاطع النصيّة Passage Retrieval، ومرحلة استخراج الإجابة Answer Extraction. حيث تركز العمل تركيزاً أساسياً على مرحلتي استخراج المقاطع النصيّة واستخراج الإجابة.

ينقسم العمل في هذا البحث إلى جزئين أساسين وهما:

مرحلة ما قبل المعالجة:

يجري في هذه المرحلة إزالة الكلمات المستبعدة، واستخراج جذوع الكلمات، ومن ثم إيجاد مرادفات الكلمات باستخدام Arabic WordNet (AWN)، تشمل هذه العمليات كلاً من كلمات السؤال المطروح وكلمات النصوص المعتمد عليها في استخراج الإجابة.

مرحلة المعالجة

يجري في هذه المرحلة قياس التشابه ما بين السؤال المطروح والمقاطع النصيّة الناتجة عن محركات البحث، حيث تم الاعتماد في قياس التشابه على الـ entailment relation واعتماداً على هذا القياس يتم إعادة ترتيب المقاطع النصيّة لاستخراج الإجابة منها.

وفيما يلي الخوارزمية المتبعة لقياس الـ entailment relation:

- Calculating the common words (c) between the question and each retrieved passage. (The common words between the question and the retrieved passage are words with the same root, and words that are related by semantic relations).
- Determining the length of each passage (m).
- Determining the length of why question (n).
- Certifying that $m \geq n \geq c$.
- Appling the three methods equations are used by [21]:

$$\cos T(T;H)=\sqrt{c/m} \quad (1)$$

$$\cos H(T;H)=\sqrt{c/n} \quad (2)$$

$$\cos H\cup T(T;H)=\sqrt{4c^2/(n+c)(m+c)} \quad (3)$$

- Satisfying this primary condition $\cos H(T,H) \geq \cos H\cup T(T,H) \geq \cos T(T,H)$
- Checking the compulsory conditions to satisfy the entailment relation. The compulsory conditions are:

$$\cos H\cup T - \cos T \leq \tau 1 \quad ...(11)$$

$$\cos H - \cos H\cup T \leq \tau \quad ...(12)$$

$$\text{Max}\{\cos T;\cos H;\cos H\cup T\} \geq \tau 3 \quad ...(13)$$

- The thresholds used in this research are: $\tau 1=0.095$, $\tau 2=0.2$, $\tau 3=0.5$.
- When all entailment conditions are checked successfully, the degree of entailment similarity which we depend on it is $\cos H\cup T$.

الشكل 14 – خوارزمية حساب entailment relation في نظام EWQA.

حيث يجري في الخوارزمية السابقة إيجاد الكلمات المشتركة بين نص السؤال والمقاطع النصية (وفقاً للتشاركها بالجذر أو في حال وجود رابط بينهما في الشبكة الدلالية)، كما يجري حساب كل من طول نص السؤال n وطول المقطع النصي m، ثم يتم حساب ثلاثة قيم تشابه ما بين نص السؤال والمقطع النصي وفقاً للمعادلات (1) و(2) و(3)، ويتم اختبار القيم المحتسبة وفقاً لمجموعة من الشروط الإلزامية (11) و(12) و(13) والتي تقوم باخبار قيم التشابه في حال تجاوزت عتبات محددة أم لا، حيث جرى تحديد قيم هذه العتبات وفقاً للتجريب والاختبار، وفي حال تحققت كافة الشروط الإلزامية يتم اعتماد قيمة التشابه النهائية (entailment similarity) على أنها القيمة المحتسبة وفقاً للمعادلة (3).

بعد استخراج المقاطع النصيّة وترتيبها يجري استخراج الجواب وفقاً للمراحل التالية:

- اختيار أول خمسة مقاطع نصيّة تم الحصول عليها من المرحلة السابقة.
- إذا كانت المقاطع النصيّة تتضمن أكثر من جملة يتم تقسيمها لعدة جمل وفقاً للنقطة.
- قياس التشابه ما بين الجمل الناتجة عن الخطوة السابقة والسؤال باعتماد طريقة الـ entailment similarity نفسها.
- اعتماد الجملة التي حصلت على أعلى قيمة تشابه كجواب عن السؤال المطروح.

- **شكل الخرج**

جملة مقتطعة من النصوص التي تشكل مجموعة بيانات النظام.

- **آلية الاختبار والنتائج**

جرى اختبار النظام من خلال طرح 250 سؤال ضمن مجالات مختلفة في (الحاسوب، الدين، العلوم، السياسة، التاريخ)، جرى وضع هذه الأسئلة من قبل ثلاثة خبراء ناطقين باللغة العربية، جرى تقييم النظام باستخدام مقياس الدقة Accuracy بحيث يعتبر الجواب صحيحاً من قبل النظام في حال كان الجواب الصحيح المرفق في عينة الاختبار ضمن أول ثلاث اقترحات يعيدها النظام، واستطاع النظام تحقيق دقة = 68.53%.

- **التكلفة**

لا يوجد في المنهجية المقترحة في هذا البحث أي عمليات تدريب، بل اعتمد بشكل أساسي على قياس تشابه النصوص وهي آليات لا تتطلب تجهيزات عتادية قوية ولا تحتاج زمن تنفيذ كبير، وبذلك لا تعتبر المنهجية المقترحة مكلفة.

2.2.2. Lemaza – 2017

- **لغة المعالجة**

يهتم النظام بالإجابة عن الأسئلة المطروحة باللغة العربية الفصحى.

- **مجال المعطيات**

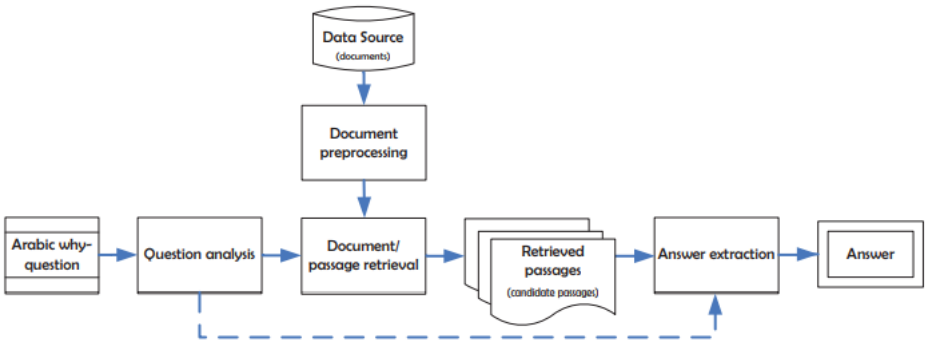
ليس هناك مجال محدد ينحصر النظام به فهو نظام إجابة آلية مفتوح المجال.

- **أنماط الأسئلة المعالجة**

يهتم النظام بالإجابة على أسئلة الاستفسار أو التعليل فقط (لماذا، ما السبب، علل وغيره).

- **آلية العمل المتّبعة**

يتألف النظام من أربعة مكوّنات أساسية [2] وهي على التتالي: مكوّن معالجة السؤال، ومكوّن استخراج المستندات، ومكوّن استخراج المقاطع النصيّة، ومكوّن استخراج الإجابة، انظر الشكل(15). يجري في مكوّن معالجة السؤال تقطيع السؤال واستخراج جذور الكلمات ومن ثم توسيع السؤال بإضافة الكلمات المرادفة (كجذور) لمفردات السؤال (معالجة نحوية ودلالية) ليصبح بذلك حقيبة من الكلمات bag of words. يجري في مكوّن استخراج المستندات معالجة نفس الخطوات المطبقة على السؤال عدا خطوة توسيع مفردات السؤال على المستندات التي تشكل مجموعة بيانات النظام واستخراج المستندات المحتمل وجود الإجابة بها وذلك باستخدام Lemur IR Toolkit[1]، ومن ثم استخراج المقاطع النصيّة في مكوّن استخراج المقاطع النصيّة من المستندات وإعطائها مرتبة (نقاط) عن طريق تطبيق استخدام نموذج Vector Space Model مع مقياس TF-IDF، يجري بعد ذلك استخراج الإجابة في المكوّن النهائي في النظام، حيث يتم ضمن هذه المرحلة استخراج الجمل الجديلة Cue Phrases وهي الجمل التي تحوي وحدات مرتبطة بأدوات ربط من نوع (بسبب، نتيجة لذلك، لام السببية، علاوة على ذلك...) ويتم تحديد الوحدة المناسبة وفقاً لنوع السؤال المطروح (تعليل، تفسير...) وإعادتها كإجابة على السؤال.



الشكل 15 – البنية العامة لنظام Lemaza.

- **شكل الخرج**

يعيد النظام جملة نصيّة مقتطعة من المستندات النصيّة كإجابة عن السؤال، أي أن الإجابة غير محددة وإنما مقطع نصي يشكل تفسيراً للسؤال المطروح.

- **آلية الاختبار والنتائج**

اعتمد النظام على مجموعة بيانات مكونة من 700 ملف ضمن مجالات مختلفة، تم الحصول عليها من Open Source Arabic Corpora(OSAC)، تم وضع مجموعة اختبار مكونة من 110 سؤالاً مع الأجوبة الموافقة لها

[1] https://sourceforge.net/projects/lemur

وذلك من قبل مجموعة أشخاص خبراء ناطقين باللغة العربية وغير منتمين لفريق البحث، وتم تقييم النظام باستخدام معياري precision و recall، واستطاع النظام تحقيق النتائج Precision = 79.2%, Recall = 72.7%.

- **التكلفة**

ليس هناك أي عمليات تدريب مكلفة، حيث اعتمد البحث بشكل أساسي على خوارزميات بسيطة غير مكلفة من ناحية الزمن والعتاد.

3.2.2. LOD – 2019

اختصاراً لـ Leveraging Linked Open Data to Automatically Answer Arabic Questions.

- **لغة المعالجة**

يهتم النظام بالإجابة عن الأسئلة المطروحة باللغة العربية الفصحى.

- **مجال المعطيات**

ليس هناك مجال محدد ينحصر النظام به فهو نظام إجابة آلية مفتوح المجال.

- **أنماط الأسئلة المعالجة**

يهتم النظام بالإجابة عن الأسئلة من الأنماط التالية:

  - الأسئلة التعريفية.
  - أسئلة التعداد.
  - أسئلة التأكيد.

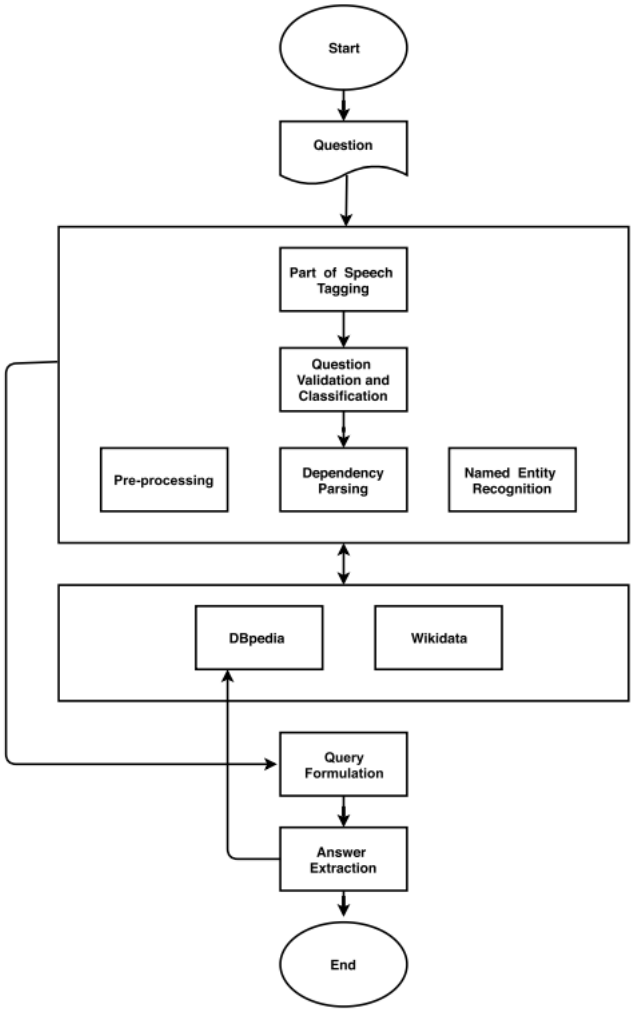
- **آلية العمل المتّبعة**

تعتمد المنهجية المقترحة في هذا البحث على ترجمة السؤال المطروح باللغة العربية (لغة طبيعية) إلى استعلام بلغة SPARQL يتم تمريره إلى قواعد معرفة مثل DBpedia[1] و Wikidata[2] لاستخراج الإجابة [6]. يمر النظام أولاً بمرحلة معالجة السؤال التي تتضمن إجراء خطوات معالجة لغوية على كلمات السؤال إضافة لاختبار صحة السؤال من ناحية الصيغة (مثال: يجب ان يبدأ بحرف استفهام وينتهي بإشارة استفهام؟)، ثم تجري مرحلة مطابقة السؤال مع ثلاثية

[1] https://wiki.dbpedia.org/

[2] https://www.wikidata.org/wiki/Wikidata:Main_Page

RDF(Resource Description Framework)[1] وذلك باستخراج العناصر الثلاثة subject و object و predicate/property من نص السؤال، وذلك بتطبيق مجموعة من القواعد لاستخراج هذه العناصر، ومن ثم تستخدم العناصر المستخرجة في تشكيل استعلام SPARQL لاستخراج الإجابة النهائية من أحد قواعد المعرفة السابقة الذكر. جرى استخدام عدة تقنيات في تحقيق الخطوات السابقة منها التعرف على الكيانات المسماة Named Entity Recognition، وإيجاد قسم الكلام Part Of Speech Tagging، و تحليل الاعتماد Dependency Parsing، ويمثل الشكل (16) البنية العامة للنظام.



الشكل 16 – البنية العامة لنظام LOD.

[1] إطار توصيف الموارد هو مجموعة من معايير رابطة الشبكة العالمية (W3C)التي صممت بداية كبيانات وصفية لنماذج بيانات. بدأ استخدامها كمنهج لعام لوصف المفاهيم، أو نمذجة المعلومات الموجودة في موارد الوب باستخدام أشكال قواعد (Syntax) متنوعة.

- **شكل الخرج**

يعيد النظام إجابة محددة وفقاً لنمط السؤال المطروح.

- **آلية الاختبار والنتائج**

جرى طرح 400 سؤال مختلف على النظام تم تجميعها بشكل يدوي، جرى تقييم النظام باستخدام معيار Precision (P) و Recall (R) و F-measure (F)، واستطاع النظام تحقيق النتائج التالية P = 84% و R = 81.3% و F = 82.8%.

- **التكلفة**

لا تحتاج الخوارزميات المذكورة في المنهجية المقترحة عتاداً خاصاً، كما لا تحتاج لتحقيقها زمناً كبيراً نسبياً، حيث لايوجد بنى شبكية وعمليات تدريب ضخمة.

4.2.2. Hybrid QAS – 2019

يهتم البحث بالإجابة على الأسئلة المطروحة باللغة العربية وذلك بالاعتماد على المعطيات المهيكلة Structured data والمعطيات النصية textual data.

- **لغة المعالجة**

يهتم النظام بالإجابة عن الأسئلة المطروحة باللغة العربية الفصحى.

- **مجال المعطيات**

لاينحصر النظام بمجال معطيات محدد فهو مفتوح المجال.

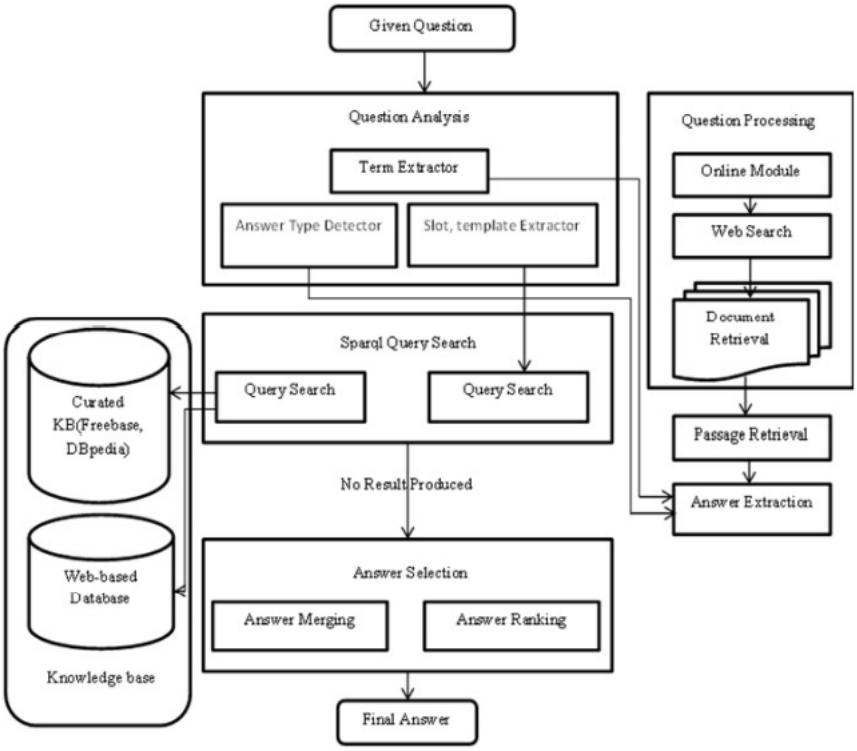
- **أنماط الأسئلة المعالجة**

ليس هناك أنماط محددة من الأسئلة التي يعالجها النظام؛ أي أن النظام غير محدد بنمط سؤال معين.

- **آلية العمل المتّبعة**

يقوم النظام باستخراج إجابات مقترحة عن الأسئلة المطروحة من مجموعة بيانات مهيكلة (قاعدة معرفة)، وكذلك استخراج إجابات مقترحة من مجموعة بيانات نصيّة (ويب) [19] ، ويعيد ترتيب هذه الإجابات وفقاً لقياس تشابه بينها وبين نص السؤال على المستوى النحوي والدلالي؛ وبذلك تجمع المنهجية المقترحة بين آلية استخراج الإجابة من البيانات المهيكلة وآلية استخراج الإجابة من البيانات النصيّة. يتألف النظام من ثلاثة أجزاء أساسية وهي: قاعدة معرفة Knowledge Base تمثل مجموعة البيانات المهيكلة التي يتم استخراج الإجابة منها، وOnline Module لاستخراج الإجابة من مجموعة البيانات النصيّة، و Text To KB Module من أجل تحويل المعطيات النصية لمعطيات مهيكلة وبالتالي إغناء مجموعة المعطيات المهيكلة المعتمدة في استخراج الإجابة (انظر الشكل(17)).

عند طرح السؤال باللغة العربية الفصحى تجري معالجته بحيث يجري تقطيعه وإزالة الكلمات المستبعدة، وتصنيفه. يتم بدايةً ترجمة السؤال إلى استعلام SPARQL لتمريره على قاعدة المعرفة المعتمدة DBpedia واستخراج مجموعة من الإجابات المقترحة له، ثم يجري استخراج مجموعة أخرى من الإجابات المقترحة باستخدام Online Module، ويجري ذلك باستخراج المقاطع النصيّة الأكثر تشابهاً مع نص السؤال وذلك من المستندات النصيّة المرجعة من البحث على الويب، ويتم جمع الإجابات المقترحة من الآليتين السابقتين ويعاد ترتيبهما وفقا لقياس التشابه النحوي والدلالي بينها وبين السؤال المقترح. جرى العمل بشكل مستقل على جزء Text To KB من أجل تحويل المعطيات النصية لمعطيات مهيكلة وبالتالي إغناء قاعدة المعرفة المعتمد عليها في استخراج الإجابة، حيث يجري تحويل المعطيات الممثلة بنصوص والتي يمكن الحصول عليها نتيجة البحث في محركات البحث التقليدية إلى معطيات مهيكلة وذلك اعتماداً على مجموعة من القوالب templates لاستخراج الثلاثيات KB Triples من الشجرة التبعية الدلالية semantic dependency tree الممثلة لجمل النصوص المرجعة من محركات البحث، حيث تم وضع مجموعة من القوالب – من قبل خبراء – توافق النماذج المختلفة لأشجار التبعية dependency trees المحتملة.



الشكل 17 – البنية العامة لنظام Hybrid QAS.

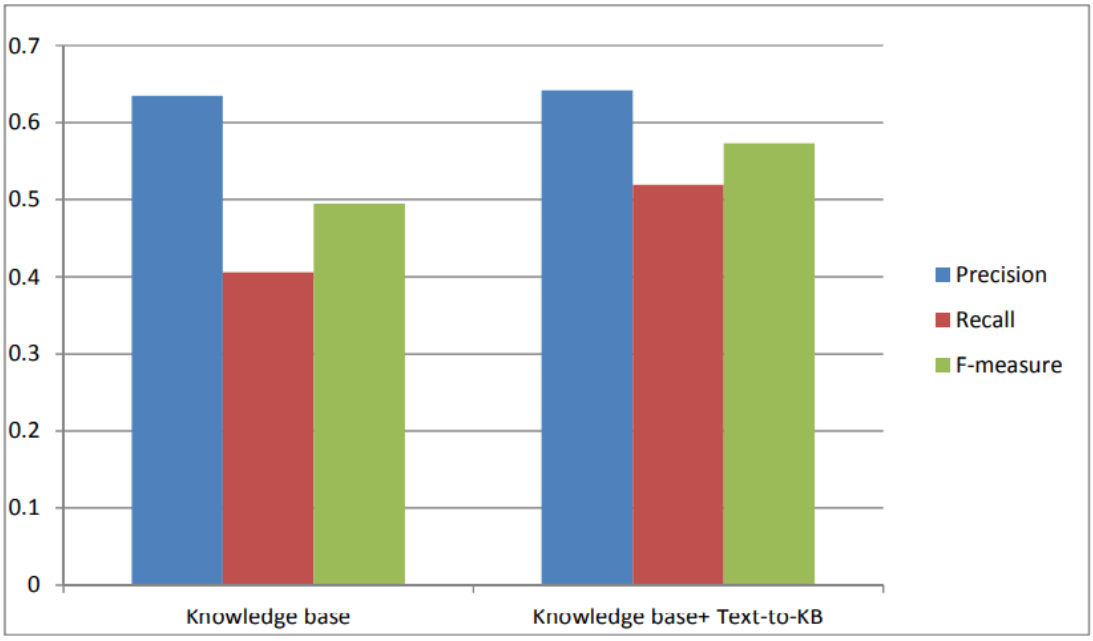
- **شكل الخرج**

لا يمكن الجزم شكل الإجابة النهائي، فيمكن للنظام أن يعيد إجابة محددة وفقاً للسؤال المطروح في حال تم اعتماد الإجابة المستخرجة من قاعدة المعرفة، اما في حال اعتماد الإجابة المستخرجة من الوب؛ فهي تشكل جزءً نصياً مقتطعاً من صفحات الويب.

- **آلية الاختبار والنتائج**

تم اختبار النظام على مجموعة بيانات منمطة مؤلفة من 1000 سؤال (لم يتم ذكر المصدر)، وجرى تقييم النظام على مرحلتين: تقييم مع استخدام مكوّن Text to KB، وتقييم دون استخدام مكوّن Text to KB ، وذلك باستخدام معايير التقويم التالية: Precession، و Recall، و F-measure، حيث استطاع النظام تحقيق النتائج المبينة في الشكل (18):

System	Precision	Recall	F1-measure
Knowledge base	.635	.406	.495
Knowledge base+ Text-to-KB (Web Search)	.642	.519	.573



الشكل 18- نتائج نظام Hybrid QAS.

- **التكلفة**

لا يوجد تتكلفة عتادية تتطلبها المنهجية المقترحة، ولكن هناك تكلفة زمنية لوضع مجموعة القواعد المعتمدة لتحويل استعلام المستخدم (السؤال) من اللغة الطبيعية للغة SPARQL، إضافة للتكلفة الزمنية لوضع القوالب المعتمدة في مكون Text to KB لتحويل النص من اللغة الطبيعية إلى ثلاثية RDF.

5.2.2. SOQAL – 2019

- **لغة المعالجة**

يهتم النظام بالإجابة عن الأسئلة المطروحة باللغة العربية الفصحى.

- **مجال المعطيات**

ليس هناك مجال محدد ينحصر النظام به فهو نظام إجابة آلية مفتوح المجال.

- **أنماط الأسئلة المعالجة**

لا يهتم النظام بأنماط محددة من الأسئلة، بحيث يجيب النظام بقطع النظر عن نمط السؤال (الإجابة غير مرتبطة بنمط السؤال).

- **آلية العمل المتّبعة**

يتألف النظام من ثلاثة مكوّنات أساسية وهي مكوّن استخراج المستندات ومكوّن استخراج الإجابات ومكوّن ترتيب الإجابات [5] (انظر الشكل(19)). دخل النظام هو السؤال المطروح باللغة العربية الفصحى والمؤلف من m كلمة ويمكن تمثيله بالمتجه $q = \{q_1, q_2, \ldots, q_m\}$، وخرج النظام هو إجابة ممثلة بجزء نصي مقتطع من مجموعة النصوص الممثلة لقاعدة بيانات النظام، وجرى هنا اعتماد مقالات الويكيبيديا كمجموعة بيانات أساسية لاقتطاع الإجابة منها.

استخراج المستندات

يجري بدايةً معالجة المستندات (مقالات الويكيبيديا) بتقطيعها، واستبدال كل كلمة بجذعها، وإزالة الكلمات المستبعدة. بعد المعالجة يجري تمثيل المستندات بمصفوفة أوزان w باعتماد مقياس TF-IDF، وكذلك يجري تمثيل السؤال q بمتجه رقمي باعتماد نفس المقياس، ويتم قياس التشابه ما بين السؤال والمستندات باستخدام قياس جيب التمام cosine similarity بين متجه السؤال ومتجه المستند، ووفقاً للقياس السابق يتم استخراج أكثر k مستند ارتباطاً بنص السؤال. يجري تقطيع المستندات الناتجة إلى فقرات ويتم استخراج أكثر p فقرة ارتباطاً بنص السؤال وفقاً لطريقة استخراج المستندات ذاتها؛ أي تم تطبيق آلية TF-IDF بشكل هرمي، أولاً لاستخراج المستندات، ثم لاستخراج الفقرات من هذه المستندات.

استخراج الإجابات

تتمثل الإجابة بجزء نصي مقتطع من الفقرات الناتجة عن المرحلة السابقة وتتعرف الإجابة بكلمة بداية i وكلمة نهاية j. جرى تحقيق مكون استخراج الإجابة باعتماد نموذج لغوي Bert Model، حيث يجري بالبداية معالجة نص السؤال والفقرات الناتجة عن المرحلة السابقة وفقاً للخطوات التالية:

- تمثيل نص السؤال بجملة وحيدة.
- تمثيل جميع الفقرات الناتجة عن المرحلة السابقة بجملة وحيدة.

- تقطيع الجملة وفقاً لطريقة Shared Workpiece[1].
- إزالة التشكيل.
- تمثيل كل من جملة السؤال وجملة الفقرات بمتجهات رقمية وذلك باستخدام آلية تضمين الكلمات المعتمد على السياق، حيث تم تدريب نموذج Bert لهذا الغرض.

جرى تخصيص نموذج Bert المدرب بإضافة طبقة إضافية دخلها نتيجة تضمين كلمات السؤال والفقرات، وخرجها متجهين رقميين الأول يمثل كلمة بداية الإجابة S وكلمة نهاية الإجابة E. يجري التدريب لكل فقرة من الفقرات الناتجة عن المرحلة السابقة بتقطيعها إلى مجموعة كلمات، ومن أجل كل كلمة i يتم حساب احتمال كون الكلمة i كلمة بداية الإجابة، واحتمال كون الكلمة i كلمة نهاية الإجابة وفقاً للمعادلتين التاليتين:

$$P_{start}(i) \propto \exp(S^T T_i)$$
$$P_{end}(i) \propto \exp(E^T T_i)$$

بحيث تمثل T_i نتيجة تضمين الكلمات للكلمة i، و(S, T) المتجهين المراد تعلمهما (الحصول عليهما) وهما متجه كلمة بداية الإجابة ومتجه كلمة نهاية الإجابة. يجري اختيار الكلمتين i و j بحيث نحصل على أكبر قيمة ممكنة للمعادلة التالية:

$$P_{start}(i) \cdot P_{end}(j)$$

مع تحقيق الشرط التالي:

$$i \leq j \leq i + 15$$

تنتج عن هذه المرحلة مجموعة مؤلفة من p إجابة مقترحة، بحيث يجري استخراج إجابة واحدة من كل فقرة من الفقرات الناتجة عن المرحلة السابقة وفقاً للطريقة المذكورة.

ترتيب الإجابات

يجري ترتيب الإجابات المقترحة الناتجة عن الخطوة السابقة بحساب مرتبة rank لكل إجابة وفقاً للمعادلة التالية:

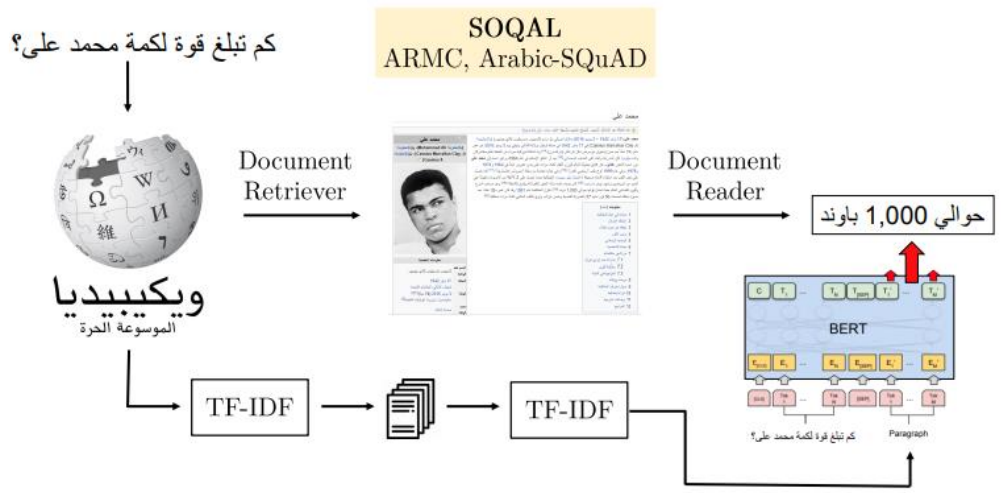
$$AnsScore(i) \propto P_{start}(i) \cdot P_{end}(i)$$

بحيث:

- AnsScore(i) تمثل قيمة تشابه الفقرة التي جرى استخراج الإجابة i منها مع نص السؤال (وفقاً لقياس التشابه المعتمد في فقرة استخراج المستندات).

[1] Shared Workpieces: وهي آلية تقطيع النص على مستوى أجزاء الكلمات وذلك باعتماد قاموس يحوي أجزاء الكلمات الشهيرة في اللغة، مثال تقطيع كلمة "قلنا" إلى "قل" و "نا".

- $P_{start}(i) \cdot P_{end}(j)$ تمثل القيمة التي تم التدريب وفقها في مرحلة استخراج الإجابات وذلك من أجل الإجابة i.



الشكل 19- البنية العامة لنظام SOQAL.

- **شكل الخرج**

مقطع نصي يشكل إجابة عن السؤال المطروح، أي أنه لا يوجد إجابة محددة وفقاً لنمط السؤال.

- **آلية الاختبار والنتائج**

تم اختبار النظام من خلال تشكيل مجموعة بيانات مستخرجة من مقالات ويكيبيديا ومؤلفة من 1395 سؤال وإجابة مع المقالات المستخرجة منها وتم نشر مجموعة البيانات المعتمدة باسم Arabic Reading Comprehension Dataset (ARCD). جرى تقييم النظام باعتماد قياس Sentence Match(SM) ما بين الإجابة الناتجة عن النظام والإجابة المرفقة في عينة الاختبار وقياس Maroco F1، واستطاع النظام تحقيق نتيجة SM = 90% و F1 = 61.3%.

- **التكلفة**

يحتاج النظام لعمليتي تدريب، عملية تدريب النموذج اللغوي Bert لإجراء خطوة تضمين الكلمات، وعملية التدريب لتخصيص النموذج السابق Bert Fine Tuning، وبذلك يحتاج النظام لعتاد قوي ولتكلفة زمنية لتحقيق عمليات التدريب المطلوبة.

6.2.2. ASHLK – 2020

- **لغة المعالجة**

يهتم النظام بالإجابة عن الأسئلة المطروحة باللغة العربية الفصحى.

- **مجال المعطيات**

يجيب النظام عن الأسئلة المطروحة في مجال الأحاديث النبويّة الشريفة والتي تشكل مجموعة بيانات النظام فهو محدد المجال.

- **أنماط الأسئلة المعالجة**

لا يهتم النظام بأنماط محددة من الأسئلة، فهو يجيب بقطع النظر عن نمط السؤال (الإجابة غير مرتبطة بنمط السؤال).

- **آلية العمل المتّبعة**

يمر النظام بثلاث مراحل أساسية [3] وهي: مرحلة المعالجة الأولية preprocessing، ومرحلة ترتيب الجمل النصية sentence ranking، ومرحلة توليد الإجابة (انظر الشكل(20)).

مرحلة المعالجة الأولية

يجري معالجة كلٍ من نص السؤال والأحاديث النبوية التي تشكل مجموعة بيانات النظام وفقاً للخطوات التالية: التقطيع، وإزالة الكلمات المستبعدة، وإزالة التشكيل، والتجذيع، والتوسيع، حيث يتم التوسيع بإغناء النص بكلمات مشابهة لكلماته، ويقاس التشابه بدرجة التشابه الدلالي بين الكلمتين، والتي يجري حسابها وفقاً للمعادلة التالية:

$$\mathrm{Sim}_{\mathrm{Dice}}(w_1, w_2) = \begin{cases} \dfrac{2 \cdot |syns(w_1) \cap syns(w_2)|}{|syns(w_1)| + |syns(w_2)|} & if\ w_1 \neq w_2 \\ 1 & if\ w_1 = w_2 \end{cases}$$

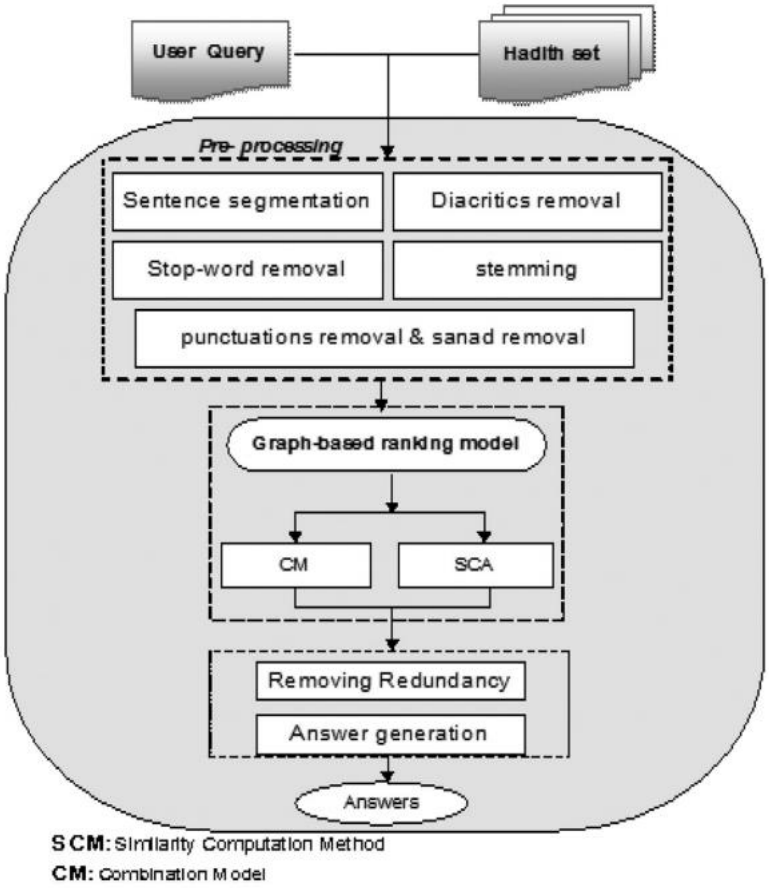
حيث syns(w) تمثل مجموعة مرادفات الكلمة w والتي يمكن الحصول عليها من أحد المعاجم الإلكترونية العربية.

مرحلة ترتيب الجمل النصية

يجري في هذا النظام تقطيع كافة الأحاديث النبوية (مجموعة البيانات) لجمل نصيّة وإعادة ترتيبها وفقاً لدرجة تشابهها مع نص السؤال، ويجري قياس التشابه وفقاً لطريقة بيانية graph based ranking، بحيث يتم تشكيل بيان كل عقدة فيه تمثل جملة نصيّة (إما من مجموعة البيانات أو الجملة التي تمثل نص السؤال)، والروابط تمثل درجة التشابه ما بين هذه الجمل (العقد). يجري قياس التشابه الكلي بقياس التشابه الدلالي Semantic Similarity ، وقياس التشابه وفقاً لترتيب الكلمات Word Order Similarity.

مرحلة توليد الإجابة

يجري توليد الإجابة باعتماد أكثر n جملةً نصيةً تشابهاً مع نص السؤال وفقاً للترتيب المطبق في المرحلة السابقة، يجري عملية حذف للجمل المتكررة وهي الجمل المتشابهة فيما بينها بشكل كبير؛ بحسب عتبة معينة يحددها مستثمرو النظام.



الشكل 20 – البنية العامة لنظام ASHLK.

- **شكل الخرج**

يعيد النظام المقاطع النصيّة المستخرجة كإجابة نهائية؛ وبالتالي الإجابة ليست محددة.

- **آلية الاختبار والنتائج**

جرى اختبار النظام على مجموعة بيانات مؤلفة من 4000 سؤال مستخرجة من 7500 حديث نبوّي. تم تقسيم الأسئلة إلى 2678 سؤال لمرحلة التدريب وتحديد معاملات النظام، 1322 سؤال مع الحديث الموافق لها لاختبار النظام. جرى تقييم النظام باستخدام المعايير الثلاثة: Precision،و Recall، وF-Measure، واستطاع النظام تحقيق النتائج التالية: Precision = 83.4%, Recall = 63.9%, F-Measure = 72.4%.

- **التكلفة**

لا تحتاج الخوارزميات المذكورة في المنهجية المقترحة عتاداً خاصاً، كما لا تحتاج لتحقيقها زمناً كبيراً نسبياً، حيث لايوجد بنى شبكية وعمليات تدريب ضخمة.

7.2.2. AraBert – 2021

لا يعد AraBert نظام إجابة آلية؛ وإنما هو نموذج لغوي BERT مدرب على مدونات باللغة العربية ومتاح للاستخدام والتخصيص في العديد من مهام معالجة اللغات الطبيعية، جرى اختبار النموذج بتخصيصه في مهمة الإجابة الآلية باللغة العربية والعديد من المهام الأخرى؛ ولكن سنهتم فقط بمهمة الإجابة الآلية للمقارنة بنظامنا لاحقاً.

- **لغة المعالجة**

يهتم النظام بالإجابة عن الأسئلة المطروحة باللغة العربية الفصحى.

- **مجال المعطيات**

يجيب النظام عن الأسئلة المطروحة في مجالات عدة؛ أي نظام إجابة آلية مفتوح المجال.

- **أنماط الأسئلة المعالجة**

لا يهتم النظام بأنماط محددة من الأسئلة، فهو يجيب بقطع النظر عن نمط السؤال.

- **آلية العمل المتّبعة**

جرى تخصيص AraBert بتطبيق نفس الآلية المتبعة في نظام SOQAL (يمكن الرجوع إليها)[15].

- **شكل الخرج**

يعيد النظام المقاطع النصيّة المستخرجة كإجابة نهائية؛ وبالتالي الإجابة ليست محددة.

- **آلية الاختبار والنتائج**

جرى اختبار النظام باستخدام Arabic Reading Comprehension Dataset (ARCD) كعينة اختبار وهي المستخدمة ذاتها في نظام SOQAL. جرى تقييم النظام باعتماد قياس Sentence Match(SM) ما بين الإجابة الناتجة عن النظام والإجابة المرفقة في عينة الاختبار وقياس Macro F1، واستطاع النظام تحقيق نتيجة SM = 92% و F1 = 62.7%.

- **التكلفة**

يحتاج النظام لعمليتي تدريب، عملية تدريب النموذج اللغوي Bert لإجراء خطوة تضمين الكلمات، وعملية التدريب لتخصيص النموذج السابق Bert Fine Tuning، وبذلك يحتاج النظام إلى عتاد قوي وإلى تكلفة زمنية لتحقيق عمليات التدريب المطلوبة.

8.2.2. AraELECTRA – 2021

وهو اختصار لـ Efficiently Learning an Encoder that Classifies Token Replacements Accurately ، ولايعد AraELECTRA نظام إجابة آلية؛ وإنما هو نموذج لغوي يشبه نموذج Bert في العمل ولكنه يختلف عنه من ناحية

البنية وآلية التدريب. جرى تدريبه على مدونات باللغة العربية ومتاح للاستخدام والتخصيص في العديد من مهام معالجة اللغات الطبيعية، جرى اختبار النموذج بتخصيصه في مهمة الإجابة الآلية باللغة العربية والعديد من المهام الأخرى؛ ولكن سنهتم فقط بمهمة الإجابة الآلية للمقارنة بنظامنا لاحقاً.

- **لغة المعالجة**

يهتم النظام بالإجابة عن الأسئلة المطروحة باللغة العربية الفصحى.

- **مجال المعطيات**

يجيب النظام عن الأسئلة المطروحة في مجالات عدة؛ أي نظام إجابة آلية مفتوح المجال.

- **أنماط الأسئلة المعالجة**

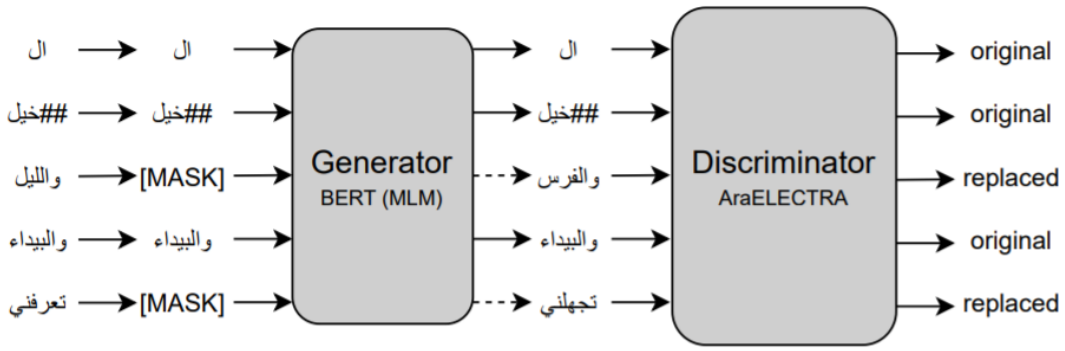
لا يهتم النظام بأنماط محددة من الأسئلة، فهو يجيب بقطع النظر عن نمط السؤال.

- **آلية العمل المتّبعة**

كما هو الحال في نماذج Bert، تجري عملية بناء النموذج اللغوي على مرحلتين وهما: مرحلة التدريب، ومرحلة التخصيص[20].

تدريب AraELECTRA

يتمثل دخل النموذج بسلسلة نصية مقطة وفقاً لطريقة التقطيع ذاتها المتبعة في تدريب نماذج Bert وهي Workpieces Tokenization، جرى تدريب النموذج باستخدام شبكتين عصبونيتين وهما: Generator(G)، وDiscriminator(D) (انظر الشكل(21)). تمثل G نموذج Bert ويجري تدريبها وفقاً لمهمة التنبؤ بالكلمات المقنعة MLM، حيث يتم تقنيع بعض كلمات الدخل ليتم التنبؤ بها بالخرج، ويدخل خرج G للشبكة التالية D والتي جرى تدريبها لكشف الكلمات المقنعة وتمييزها عن الكلمات الأصلية في الجملة كما هو موضح بالشكل(21).



الشكل 21 – البنية العامة لنموذج AraELECTRA.

تخصيص AraELECTRA

جرى تخصيص AraELECTRA بتطبيق نفس الآلية المتبعة في نظام AraBert.

- **شكل الخرج**

يعيد النظام المقاطع النصيّة المستخرجة كإجابة نهائية؛ وبالتالي الإجابة ليست محددة.

- **آلية الاختبار والنتائج**

جرى اختبار النظام باستخدام Arabic Reading Comprehension Dataset (ARCD) كعينة اختبار وهي المستخدمة ذاتها في AraBert. جرى تقييم النظام باعتماد قياس Exact Match(EM) ما بين الإجابة الناتجة عن النظام والإجابة المرفقة في عينة الاختبار و F-Measure، واستطاع النظام تحقيق نتيجة EM = 37% و F1 = 71.22%

- **التكلفة**

يحتاج النظام لعمليتي تدريب، عملية تدريب النموذج اللغوي لإجراء خطوة تضمين الكلمات، وعملية التدريب لتخصيص النموذج السابق Fine Tuning، وبذلك يحتاج النظام لعتاد قوي ولتكلفة زمنية لتحقيق عمليات التدريب المطلوبة.

3.مقارنة

1.3.نقاط القوة في الأعمال السابقة

- وجود خوارزمية فعّالة في بعض نظم الإجابة الآلية باللغة الإنكليزية قادرة على استخراج إجابات محددة وفقاً لنمط السؤال المطروح، كما في SimBioNLQA.
- تطبيق آلية معالجة نحوية ودلالية في قياس التشابه بين النصوص وذلك في مرحلتي استخراج المقاطع النصية واستخراج الإجابة مما يزيد من دقة النظام.
- استخدام النماذج اللغوية في استخراج الإجابة وذلك في النظم العربية والأجنبية؛ وقد استطاعت تحقيق نتائج جيدة جداً في ذلك.

2.3.نقاط الضعف في الأعمال السابقة

- اقتصار بعض نظم الإجابة الآلية على الإجابة عن أسئلة ضمن مجال محدد، كما في نظم SimBioNLQA و ASHLK.
- اقتصار بعض نظم الإجابة الآلية العربية على الإجابة عن أسئلة من نمط محدد، كما في Lemaza و EWAQ.
- اقتصار بعض نظم الإجابة الآلية العربية على الإجابة عن الأسئلة المطروحة بمقطع نصي مقتطع من المستندات النصية التي تشكل مجموعة بيانات النظام بدلاً من تقديم إجابة دقيقة ومحددة وفقاً لنمط السؤال المطروح.

- التكلفة المادية (العتاد) والزمنية لتحقيق النماذج اللغوية التي اعتمدت عليها أغلب الدراسات السابقة الحديثة.

3.3.الخلاصة

بعد استعراض النظم السابقة نلاحظ أن نظام SemBioNLQA هو النظام الأكثر شمولية بحيث أنه يعالج أنماطاً مختلفة من الأسئلة ويعيد إجابات محددة ودقيقة من أجل كل نمط، كما أنه يعتمد في استخراج الإجابات على المعالجة النحوية والدلالية، إلا أنه يعالج اللغة الإنكليزية ويقتصر على مجال محدد. أما بالنسبة للنظم التي تُعنى باللغة العربية، فنلاحظ أن النظم المعتمدة على النماذج اللغوية، كما في SOQAL و AraBert و AraELECTRA استطاعت تحقيق أفضل النتائج مقارنة بغيرها بالرغم من معالجتها لأنماط مختلفة من الأسئلة وضمن مجالات مفتوحة، لكنها تقتصر على إعادة مقاطع نصية كإجابات نهائية ولا تهتم باستخراج الإجابة المحددة وفقاً لنمط السؤال. يوضح الجدول (4) مقارنةً للدراسات السابقة المشابهة وفقاً للمعايير المختلفة التي تم تحديدها في بداية الفصل.

بناءً على ما سبق تعتبر الدراسات حول نظم الإجابة الآلية باللغة العربية جيدة إلا أنها مقيدة بعض الشي ولا ترقى للنظم المماثلة في اللغات الأخرى في الحصول على إجابات دقيقة، وبذلك يمكن لنا العمل على هذه الثغرة في بحثنا، بحيث يمكننا الاستفادة من الدراسات حول نظم اللغة العربية مثل SOQAL في تحقيق مكوّنات استخراج المستندات والمقاطع النصيّة ذات الصلة بالسؤال المطروح وذلك لما حققته من نتائج جيدة ضمن هذا المجال، كما يمكن الاستفادة من دراسات النظم في اللغات الإنكليزية مثل SemBioNLAQ في تحقيق مكوّن استخراج الإجابة الذي يعيد إجابات محددة ودقيقة وفقاً لنمط السؤال المطروح.

معيار المقارنة/النظام	لغة المعالجة	مجال المعطيات	أنماط الأسئلة المعالجة	شكل الخرج	آلية الاختبار والنتائج							التكلفة
					مجموعة البيانات	P	R	F1	EM	SM	Acc	
DrQA	الإنكليزية	مفتوحة المجال	محددة النمط	مقطع نصي	ثلاث مجموعات بيانات منمطة	-	-	79%	70%	-	-	مكلف
Efficient QA	الإنكليزية والفرنسية	مفتوحة المجال	غير محددة النمط	مقطع نصي	مجموعة بيانات منمّطة غير محددة الحجم	-	-	96.7%	92.3%	-	-	مكلف
SemBioNLQA	الإنكليزية	محددة المجال	غير محدد النمط	إجابة محددة	مجموعة بيانات منمطة	موضحة بالتفصيل بالجدول رقم (3)					-	غير مكلف
Lemaza	العربية	مفتوحة المجال	محددة بنمط وحيد	مقطع نصي	110سؤال	79.2%	72.2%	-	-	-	-	غير مكلف
EWAQ	العربية	مفتوحة المجال	محددة بنمط وحيد	مقطع نصي	250 سؤال	-	-	-	-	-	68.3%	غير مكلف
LOD	العربية	مفتوحة المجال	محددة بعدة أنماط	إجابة محددة	400 سؤال	84%	81.3%	82.8%	-	-	-	غير مكلف
Hybrid QAS	العربية	مفتوحة المجال	غير محددة النمط	-	1000سؤال	64.2%	51.9%	57.3%	-	-	-	غير مكلف
SOQAL	العربية الفصحى	مفتوحة المجال	غير محددة النمط	مقطع نصي	ACRD	-	-	61.3%	-	90%	-	مكلف
ASHLK	العربية الفصحى	محددة المجال	غير محددة النمط	مقطع نصي	1322سؤال	83.4%	63.9%	72.4%	-	-	-	غير مكلف
AraBert	العربية الفصحى	مفتوحة المجال	غير محددة النمط	مقطع نصي	ACRD	-	-	62.7%	-	-	92%	مكلف
AraELECTRA	العربية الفصحى	مفتوحة المجال	غير محددة النمط	مقطع نصي	ACRD	-	-	71.2%	37%	-	-	مكلف

الجدول 4 – مقارنة للدراسات السابقة المشابهة.

4.الخاتمة

جرى في هذا الفصل عرض أهم الدراسات والأبحاث المشابهة لبحثنا والتي أجريت في السنوات الأخيرة، بحيث تم استعراض كل بحث من حيث مجموعة بيانته المعتمدة وآلية عمله والنتائج الحاصل عليها وتكلفة تنفيذه، كما جرى مقارنة هذه الأبحاث وفقاً لعدة معايير قمنا بتحديدها في بداية الفصل. وبناءً على هذه الدراسة تم اعتماد منهجية العمل التي سيرتكز عليها بحثنا.

الفصل الرابع: المنهجية المقترحة

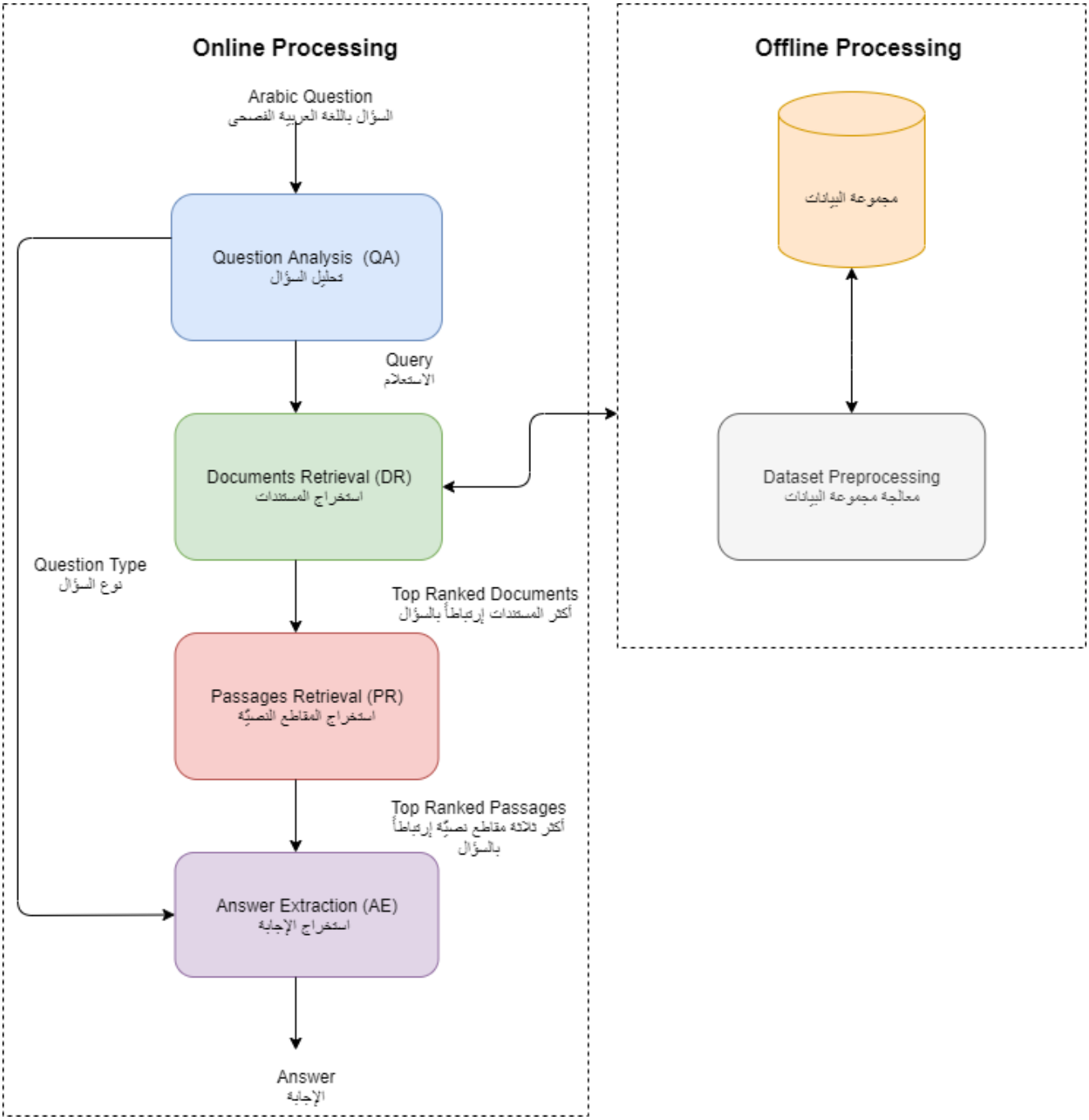
تعتمد المنهجية المقترحة على إيجاد نظام إجابة آلية مشابه في بنيته العامة للنظم السابقة، والمؤلفة من المكوّنات الأربعة وهي: معالجة السؤال، واستخراج المستندات، واستخراج المقاطع النصيّة، واستخراج الإجابة، بحيث يكون دخل النظام سؤالاً مطروحاً باللغة العربية الفصحى وغير محصور بمجال محدد (نظام إجابة آلية مفتوح المجال)، ويمكنه معالجة أسئلة بأنماط مختلفة (غير محصور بنمط أسئلة محدد). ويتمثل الخرج بإجابة دقيقة وواضحة وفقاً لنمط السؤال المطروح.

نهدف أن تعالج المنهجية المقترحة الثغرات الموجودة في النظم السابقة قدر الإمكان؛ وفي سبيل تحقيق ذلك ارتكز بحثنا على مرحلة استخراج المقاطع النصيّة، ويعود ذلك لتأثيرها الكبير على الخرج النهائي للنظام. فالإجابة الدقيقة المراد الحصول عليها يتم استخراجها من مقطع نصي يفترض أن يكون صحيحاً. وتم تحقيق ذلك من خلال تطبيق تقنيات حديثة في مجال معالجة اللغات الطبيعة تأخذ بالاعتبار المعالجة على المستوى النحوي والدلالي، وكذلك اهتم بحثنا ايضاً بمرحلة استخراج الإجابة، حيث جرى إيجاد وتطبيق خوارزمية فعّالة في مرحلة استخراج الإجابة مستوحاة من خوارزمية مطبقة في نظم الإجابة الآلية باللغة الإنكليزية [7]، ولم يسبق تطبيقها في النظم المشابهة السابقة بالنسبة للغة العربية، تعتمد هذه الخوارزمية على إجراء معالجة لغوية نحوياً ودلالياً على المقاطع النصيّة الناتجة عن مرحلة استخراج المقاطع النصيّة وذلك وفقاً لنمط السؤال المطروح، بهدف الحصول على إجابة دقيقة ومحددة.

1.المخطط العام للمنهجية المقترحة

يتألف النظام من مرحلتي معالجة رئيستين وهما: مرحلة ما قبل التشغيل (Offline Processing) ومرحلة التشغيل (Online Processing)، حيث يجري في مرحلة ما قبل التشغيل معالجة مجموعة البيانات (النصوص) المعتمدة وإعادة صياغتها بشكل يمكن التعامل معه لاحقاً في مراحل المعالجة أثناء التشغيل(انظر الشكل (22))، أما في مرحلة التشغيل وهي المرحلة التي يصبح بها النظام قابل للاستثمار من قبل المستخدمين، يتمثل دخل النظام بسؤال مطروح باللغة العربية الفصحى، وتتم معالجته مروراً بأربعة مراحل وهي على التتالي: مرحلة تحليل السؤال، والتي تهتم بمعالجته لغوياً وتصنيفه لأحد الأنماط التي يعتمدها النظام وإعادة صياغته كاستعلام لمرحلة استخراج المستندات، والتي تهتم باستخراج أكثر المستندات الأكثر ارتباطاً بهذا

الاستعلام وفقاً لمعايير نحوية ودلالية محددة، وتكوّن هذه المستندات دخلاً لمرحلة استخراج المقاطع النصيّة، والتي تستخرج مقاطع (جملاً) تمثل أجزاءً من هذه المستندات ومتوقع ورود الإجابة ضمنها، تدخل المقاطع النصيّة مرحلة استخراج الإجابة، ويتم إجراء المعالجة اللازمة عليها وفقاً لنمط السؤال الذي يتم تحديده في مرحلة تحليل السؤال، وتنتج عن هذه المرحلة إجابة مختصرة وواضحة ودقيقة موافقة لنمط السؤال المطروح وتمثل الخرج النهائي للنظام.



الشكل 22 – البنية العامة للنظام المقترح.

1.1.مرحلة ما قبل التشغيل Offline

يجري بالبداية معالجة كافة المستندات التي تمثل مجموعة بيانات النظام، بحيث قمنا بإجراء خطوات المعالجة التالية:

- تقطيع نص المستند إلى كلمات وجرى التقطيع وفقاً للفراغ وعلامات الترقيم.
- إزالة الأحرف والكلمات غير المنتمية لأبجدية اللغة العربية.
- إزالة علامات الترقيم.
- إزالة الأرقام.
- استنظام النص[1].
- تجذيع كلمات النص (باستبدال كل كلمة بالجذع).
- إزالة كلمات الوقف[2].

جرى تحقيق آلية البحث عن المستندات باستخدام نموذج فضاء المتجه[3] Vector Space Model (VSM) ، بحيث يتمثل كل مستند بمتجه رقمي، وبُعد هذا المتجه مساوياً لعدد المصطلحات Terms (T) المذكورة في كافة المستندات دون تكرار (في حال كان عدد المصطلحات في المستندات التي تمثل مجموعة بيانات النظام مساوياً لـ n فيكون بُعد المتجه الواحد هو n)، وجرى حساب قيم المتجهات اعتماداً على قياس TF-IDF، فإذاكان لدينا D مستند، سينتج عن هذه المرحلة مجموعة متجهات S_D مؤلفة من D متجهاً من البعد n يمكن تمثيلها بالشكل التالي:

$$S_D = \{d_1, d_2, d_3, \ldots, d_i, \ldots, d_D\}$$

$$d_i = [\text{TF-IDF}(S_D, d_i, T_1), \text{TF-IDF}(S_D, d_i, T_2), \ldots, \text{TF-IDF}(S_D, d_i, T_n)]$$

بحيث يتم حساب قياس TF-IDF وفقاً للمعادلات التالية [21]:

$$TF - IDF(SD, d, T) = TF(T, d).IDF(T, SD)$$

$$TF(T, d) = \log\big(1 + frequency(T, d)\big)$$

$$IDF(T, SD) = \log\left(\frac{n}{count\ (d\ \in SD\ ; T\ \in d)}\right)$$

---

[1] **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").

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

[3] **نموذج فضاء المتجه** أو نموذج متجه المصطلحات: 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 إحدى الطرائق المعتمدة لحساب قيم هذه المتجهات TF-IDF. يستخدم هذا النموذج في نظم استرجاع المعلومات والفهرسة والترتيب الإرتباطي.

2.1.مرحلة التشغيل Online

وهي المرحلة التي يوضع بها النظام للاستثمار من قبل المستخدمين، بحيث يقوم المستخدم بطرح السؤال والحصول على الإجابة، وتجري معالجة السؤال وفقاً للمراحل التالية (**في كل مرحلة من المراحل التالية هناك قيمة مضافة لبحثنا وسنقوم بتوضيح ذلك من أجل كل مرحلة**):

1.2.1.المرحلة الأولى: تحليل السؤال

يجري في هذه المرحلة إجراء المعالجة اللازمة على نص السؤال المطروح وفقاً لعدة خطوات وهي: معالجة السؤال لغوياً، وتصنيف السؤال، وتوسيع السؤال (انظر الشكل (22)).

1.1.2.1.معالجة السؤال لغوياً

طبقنا في هذه المرحلة مجموعة من الإجراءات على كلمات السؤال ليصبح بصيغة استعلام يمكن التعامل معه في المراحل اللاحقة، ويمكن تلخيص هذه الإجراءات بالخطوات التالية:

- تقطيع السؤال إلى كلمات وجرى التقطيع وفقاً للفراغ وعلامات الترقيم.
- إزالة الأحرف والكلمات غير المنتمية لأبجدية اللغة العربية.
- إزالة علامات الترقيم.
- إزالة الأرقام.
- استنظام نص السؤال[1].
- تجذيع كلمات السؤال (باستبدال كل كلمة بالجذع).

تم تحديد المراحل السابقة واعتمادها وفقاً لاختبار دقة النظام عند كل مرحلة وسنبيّن ذلك بالتفصيل في الفصل التالي. استخدمنا مكتبة NLTK[2] لتحقيق خطوات المعالجة السابقة.

---

[1] **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").

[2] https://www.nltk.org/data.html

2.1.2.1.تصنيف السؤال

قمنا في هذه المرحلة بتصنيف السؤال لأحد الأنماط الرئيسية التي يعالجها النظام، وذلك بغرض استخراج الإجابة الدقيقة لاحقاً في مرحلة استخراج الإجابة وفقاً لنمط السؤال المطروح. يبين الجدول[1](5) أنماط الأسئلة التي يعالجها بحثنا.

نمط السؤال	مثال
الأسئلة التعريفية Factoid Questions	"أين عاش أحمد عمر خاشقجي؟"
أسئلة التعداد List Questions	"ما هي المهن التي عمل بها خاشقجي؟"
أسئلة التأكيد Confirmation Questions	"هل قرة القدم رياضة شعبية؟"
أسئلة التلخيص Summarization Questions	"لماذا سميت غزوة بدر بهذا الاسم؟"

الجدول 5 – أنماط الأسئلة التي يعالجها البحث.

تم تصنيف الأسئلة باستخدام مصنف Support Vector Machine (SVM)، بحيث دخل المصنف هو السؤال بعد أن أُجريت عليه خطوات المعالجة المذكورة سابقاً، وجرى استخدام نموذج حقيبة الكلمات Bag of Words[2] (BOW) إلى جانب قياس TF-IDF[3] كميزة لتدريب المصنف، وجرى اعتماد المصنف والميزات المستخدمة له وفقاً لاختبار دقة النظام عند عدة مصنفات وميزات أخرى سيتم ذكرها بالتفصيل في الفصل التالي. جرى تحقيق المصنف السابق باستخدام الأدوات المتاحة في مكتبة Scikit-learn[4].

---

[1] يوضح الجدول (2) في الفصل الثالث من البحث شرحاً للأنماط المذكورة.

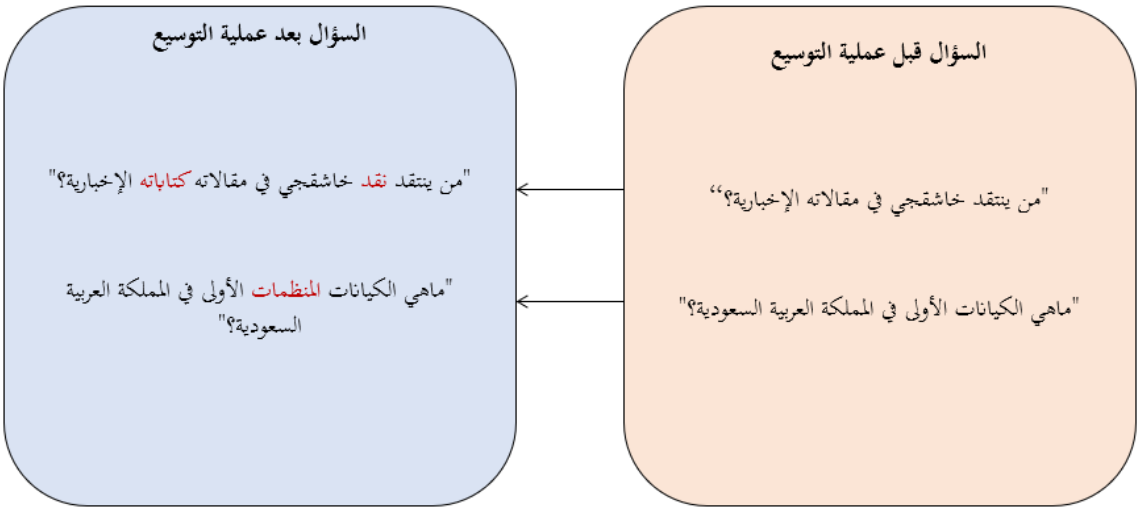
[2] **نموذج حقيبة الكلمات**: هو تمثيل مبسط يستخدم في معالجة اللغة الطبيعية واسترجاع المعلومات. يتم فيه تمثيل النصوص كالجمل أو المستندات كحقيبة (مجموعة متعددة) من الكلمات الواردة فيها، متجاهلاً القواعد اللغوية وترتيب الكلمات مع الحفاظ على التعددية. يستخدم نموذج حقيبة الكلمات بشكل شائع في تصنيف النصوص حيث يتم استخدام (تكرارات) كل كلمة كميزة لتدريب المصنف.

[3] **TF-IDF**: هو مقياس إحصائي يقيم مدى صلة كلمة ما بوثيقة في مجموعة من الوثائق. يتم ذلك بضرب مقياسين: عدد المرات التي تظهر فيها الكلمة في المستند ، وتواتر المستند العكسي للكلمة عبر مجموعة من المستندات.

[4] https://scikit-learn.org/stable/install.html

3.1.2.1.توسيع السؤال

قمنا في هذه المرحلة بتطبيق تقنية توسيع الاستعلام[1]، وذلك بهدف زيادة دقة عملية استرجاع النصوص في كل من مرحلتي استخراج المستندات واستخراج المقاطع النصيّة. جرى تحقيق ذلك بأخذ مرادف وحيد (إن وجد) لكل كلمة من كلمات السؤال وإضافته لنص السؤال, كذلك، جرى استخراج المرادفات باستخدام تقنية تضمين الكلمات المستقلة عن السياق (نموذج Word2Vec) **وهنا تكمن القيمة المضافة لبحثنا بالنسبة لمرحلة توسيع السؤال حيث لم يسبق استخدام مفهوم تضمين الكلمات في مرحلة توسيع السؤال من قبل**، وجرى الاعتماد على نموذج تضمين كلمات مدرب مسبقاً ([12] AraVec) يأخذ الكلمة ويعيد أكثر كلمة متشابهة معها في السياق (تردان في سياق واحد عادةً). يوضح الشكل(23) بعض الأمثلة لعملية توسيع السؤال باستخدام نموذج تضمين الكلمات AraVec.



الشكل 23 – أمثلة عن توسيع السؤال باستخدام نموذج AraVec.

نلاحظ من الأمثلة السابقة بأنه لا تؤخذ مرادفات كافة كلمات السؤال، ويعود ذلك لنموذج تضمين الكلمات المستخدم، والمدونات المدرب عليها (قد لا تحوي مدونات التدريب على الكلمة وبالتالي لايمكن استخراج مرادفها).

جرى اعتماد آلية توسيع السؤال بالإضافة لعدد المرادفات المستخدمة لتوسيعه وفقاً لاختبار دقة النظام باستعمال عدة آليات تُستخدم عادةً لهذا الغرض (مثل الشبكات الدلالية وغيرها)، ومن أجل عدة مرادفات. وسنبين ذلك بالتفصيل في الفصل التالي.

---

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

2.2.1.المرحلة الثانية: استخراج المستندات

قمنا في هذه المرحلة باستخراج مجموعة المستندات المرتبطة بالسؤال، بحيث دخل هذه المرحلة هو السؤال الناتج عن المرحلة السابقة (السؤال بعد المعالجة والتوسيع)، على شكل استعلام، للبحث عن أكثر المستندات ارتباطاً به. تجري مرحلة استخراج المستندات وفقاً للخطوات التالية:

1.2.2.1.استخراج المستندات المرتبطة بالسؤال

يتم تمثيل نص السؤال وفقاً لطريقة تمثيل المستندات (مرحلة ما قبل التشغيل)، بحيث يكون متجه السؤال كالتالي:

$$q = [\text{TF-IDF}\ (S_D, q, T_1), \text{TF-IDF}\ (S_D, q, T_2), \ldots, \text{TF-IDF}\ (S_D, q, T_n)]$$

قمنا بقياس التشابه ما بين المستندات النصيّة ونص السؤال باستخدام قياس جيب التمام Cosine Similarity ما بين متجهات المستندات ومتجه السؤال، والذي يتم حسابه وفقاً للمعادلة التالية [22] :

$$Cosine\ Similarity(d, q) = \frac{d.q}{||d||.||q||}$$

2.2.2.1.ترتيب المستندات

بعد أن قمنا بحساب تشابه نص السؤال مع كل مستند من المستندات وفقاً للقياس المذكور سابقاً، جرى ترتيب هذه المستندات وفقاً لقيمة التشابه من الأعلى للأدنى، واخترنا أكثر 5 مستندات تشابهاً مع نص السؤال لتدخل للمرحلة اللاحقة (مرحلة استخراج المقاطع النصيّة).

3.2.1.المرحلة الثالثة: استخراج المقاطع النصيّة

يجري في هذه المرحلة استخراج أجزاء نصيّة من المستندات الناتجة عن المرحلة السابقة، بحيث اعتبرنا المقطع النصي هو الجزء من النص الذي ينتهي بنقطة.

كما يوضح الشكل(22)، تجري هذه المرحلة وفقاً للخطوات التالية:

1.3.2.1.استخراج المقاطع النصيّة من المستندات

جرت معالجة المستندات وفقاً لمراحل المعالجة اللغوية المذكورة سابقاً، إضافة لمرحلة التقطيع وفقاً للنقطة "." (أي تقطيع المستندات وفقاً لوجود النقطة)، ونتج بذلك مجموعة من المقاطع النصيّة.

2.3.2.1.ترتيب المقاطع النصيّة

قمنا بترتيب المقاطع النصيّة المستخرجة وفقاً لمقدار التشابه ما بين المقطع النصي ونص السؤال، وقمنا بقياس التشابه اعتماداً على قياسين نحوي ودلالي كما يلي:

قياس التشابه نحوياً

يجري قياس التشابه ما بين نص السؤال (بعد المعالجة) والمقاطع النصية المستخرجة استناداً لتشابههما نحوياً. جرى قياس التشابه باستخدام تابع BM25[1] الشهير، يأخذ كدخل جملتين باللغة الطبيعية ويعيد قيمة تعبر عن مدى تقارب هاتين الجملتين من الناحية المعجمية (مدى احتوائهما على كلمات مشتركة)، حيث يجري حساب التشابه في تابع BM25 وفقاً للمعادلة التالية:

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

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 (شرحنا كيفية حسابه في فقرة استخراج المستندات).

قياس التشابه دلالياً

قمنا بدعم القياس السابق بإضافة قياس جديد يأخذ بالاعتبار التشابه الدلالي ما بين جملتين، وذلك من أجل زيادة دقة النظام واستبعاد الحالات غير الدقيقة. ففي الكثير من الأحيان يمكن أن ترد جملتان متقاربتان بشكل كبير، ولكن لا يمكن اكتشاف ذلك باستخدام المقياس السابق لوحده، مثال: جملة "جمال أحمد خاشقجي، صحفي وإعلامي سعودي." وجملة "جمال أحمد خاشقجي، رأس عدّة مناصب لعدد من الصحف في السعودية." نلاحظ أن الجملتين تشتركان فقط بـ "جمال أحمد خاشقجي" على الرغم من مدى التقارب الشديد بينهما دلالياً. اعتمدنا في قياس التشابه الدلالي على تقنية تضمين الكلمات المستقلة عن

[1] **BM25** (**Okapi BM25**): اختصاراً لكلمة أفضل مطابقة Best Matching، هو تابع ترتيب تستخدمه محركات البحث لتقدير مدى صلة المستندات باستعلام بحث معين.

السياق، حيث قمنا باستخدام نموذج تضمين كلمات مدرب مسبقاً AraVec[1] (وهو نفسه المستخدم في مرحلة توسيع السؤال)، بحيث جرى تمثيل كل كلمة بمتجه رقمي استناداً لنموذج AraVec، وجرى قياس تشابه الكلمات باستخدام مقياس جيب التمام الذي قمنا بذكره سابقاً ما بين هذه المتجهات. يوضح الشكل (24) الخوارزمية التي قمنا بوضعها لحساب التشابه الدلالي وذلك بالاعتماد على نموذج تضمين الكلمات AraVec:

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

الخرج: قيمة التشابه بين S_1 و S_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}:
   a. من أجل كل كلمة w_2 في القائمة S_{words2}:
      i. استخراج متجه الكلمة w1 الناتج عن نموذج تضمين الكلمات ووضعه في vec_1.
      ii. استخراج متجه الكلمة w2 الناتج عن نموذج تضمين الكلمات ووضعه في vec_2.
      iii. 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.

الشكل 24 – خوارزمية حساب التشابه باعتماد نموذج تضمين الكلمات AraVec.

قمنا بتطبيق الخوارزمية السابقة من أجل كل مقطع نصي مع نص السؤال؛ أي في حال نتج لدينا من تقطيع المستندات الخمسة مجموعة من المقاطع النصيّة S_P مؤلفة من P مقطعاً نصياً كالتالي:

$$S_P = \{p_1, p_2, p_3, \ldots, p_P\}$$

نقوم بحساب التشابه الدلالي ما بين السؤال q وكل مقطع من المقاطع السابقة.

قياس التشابه الكلي

قمنا بوضع نموذج خطي يجمع ما بين القياس النحوي والدلالي السابقين وذلك لحساب التشابه النهائي ما بين نص السؤال q والمقطع النصي p ، ويمكن تمثيله بالشكل التالي:

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

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

جرى تحديد قيمة المعامل α تجريبياً كما سنوضح في الفصل التالي. كما جرى ترتيب المقاطع النصيّة وفقاً لقياس التشابه TotalSim السابق من الأعلى للأدنى، واخترنا أكثر 3 مقاطع تشابهاً مع نص السؤال لتدخل للمرحلة اللاحقة (مرحة استخراج الإجابة).

إن استخدام مفهوم تضمين الكلمات لقياس التشابه الدلالي، والجمع ما بين التشابه الدلالي والنحوي وفقاً لنموذج خطي مع استخدام معامل توزين يمثل القيمة المضافة لبحثنا في مرحلة استخراج المقاطع النصية.

4.2.1.المرحلة الرابعة: مرحلة استخراج الإجابة

تختلف آلية استخراج الإجابة باختلاف نمط السؤال المطروح، وبذلك لدينا أربع آليات جرى اتباعها في استخراج الإجابة، وفقاً لكل نمط من أنماط الأسئلة التي يعالجها النظام. قمنا بوضع الآليات المختلفة لاستخراج الإجابة استناداً لخوارزميات مشابهة تم تطبيقها في نظام إجابة آلية باللغة الإنكليزية [7].

1.4.2.1.الإجابة على أسئلة التأكيد

تجري، في نظامنا، الإجابة على هذا النمط من الأسئلة بكلمة "نعم" أو "لا". اعتمدنا في تحقيق ذلك على تقنية تحليل المشاعر[1] Sentiment Analysis، بحيث يجري تصنيف المقاطع النصية الناتجة عن المرحلة السابقة P بحسب المشاعر التي تعبر عنها إلى صنفين "إيجابية" أو "سلبية"، ويجري كذلك تصنيف السؤال بحسب المشاعر التي يعبر عنها، ويجيب النظام بـ"نعم" أو "لا" تبعاً لصنف المشاعر في كلاً من المقاطع النصيّة والسؤال؛ فإذا توافقت المشاعر في إيجابيتها أو سلبيتها في كل من السؤال والمقاطع النصية معاً يجيب النظام بـ"نعم" وإلا يجيب النظام بـ "لا". يوضح الشكل(25) الخوارزمية التي قمنا بوضعها للإجابة على هذا النمط من الأسئلة.

[1] **تحليل المشاعر أو الآراء**: هو استخدام معالجة اللغات الطبيعية، وعلم اللغة الحاسوبي والتحليل النصي من أجل الكشف عما يحمله النص من مشاعر سواء إيجابية أو سلبية أو محايدة تجاه موضوع النص.

الدخل: قائمة بثلاث مقاطع نصيّة (الناتجة عن مرحلة استخراج المقاطع النصيّة) S_p، ونص السؤال q.

الخرج: "نعم" أو "لا".

الخطوات:

1. تحليل المشاعر في نص السؤال ووضع النتيجة في $Sentiment_{question}$.
2. تهيئة المتحولات $Score_{Positive} = 0$ و $Score_{Negative} = 0$.
3. من أجل كل مقطع نصي p من قائمة المقاطع النصية S_p:
   a. تحليل المشاعر للمقطع النصي p ووضع النتيجة في $Sentiment_p$.
   b. إذا كانت $Sentiment_p$ = 'negative':

      $Score_{Negative} = Score_{Negative} + 1$.

   c. وإلا:

      $Score_{Positive} = Score_{Positive} + 1$.
4. إذا كان $Score_{Positive} > Score_{Negative}$.

   $Sentiment_{Passages}$ = 'positive'

5. وإلا:

   $Sentiment_{Passages}$ = 'negative

6. إذا كان $Sentiment_{question} == Sentiment_{Passages}$.

   الإجابة بـ "نعم"

7. وإلا:

   الإجابة بـ "لا"

الشكل 25 – خوارزمية الإجابة على أسئلة التأكيد.

جرى وضع الخوارزمية السابقة استناداً إلى الخوارزمية المطبقة في نظام SimBioNLQA[7] **مع إضافة خطوة تحليل المشاعر في نص السؤال والتحقق من توافق مشاعر السؤال مع المقاطع النصية وهنا تكمن إضافتنا، حيث اكتفت الخوارزمية**

المطروحة في بحث SimbioNLQA بتحليل مشاعر المقاطع النصية فقط واعتمادها في الإجابة النهائية. استخدمنا في تحقيق مرحلة تحليل المشاعر أداة مزاجك[1] mazajak، وهي أداة مجانية تستخدم لتحليل المشاعر في النصوص العربية [23].

2.4.2.1.الإجابة عن الأسئلة التعريفية

اعتمدنا في الإجابة على هذا النمط من الأسئلة على تقنية التعرف على الكيانات المسمّاة[2]، بحيث جرى في البداية تصنيف السؤال التعريفي لأحد الأصناف الثلاث التالية: سؤال عن مكان، أو عن أشخاص، أو عن منظمة ما. واستخراج الكيان الموافق تبعاً لصنف السؤال. ولكن قبل استخراج الكيانات المسماة من المقاطع النصيّة؛ كان لابد من إجراء خطوة إضافية للتقليل من الاحتمالات الكثيرة الممكن ورودها في المقاطع النصيّة (**وهنا تكمن إضافتنا**)، مثال: لدينا السؤال "أين ولد رسول الله؟"، وهو سؤال عن مكان، والمقطع النصي "ولد رسول الله(ص) في مكة المكرمة وتوفي في المدينة المنورة"، لدينا احتمالان للإجابة وهما "مكة المكرمة" و"المدينة المنورة". إذن، نحن بحاجة خطوة إضافية تقوم باختصار المقطع النصي عن طريق اقتطاع جزء قصير منه ويحمل الاحتمال الأكبر لورود الإجابة الصحيحة ضمنه. جرى تحقيق هذه الخطوة وفقاً للمنهجية المتبعة في نظام SOQAL، والتي اعتمدت على نموذج لغوي مدرب مسبقاً BERT لإجراء خطوة تضمين الكلمات، ومن ثم تخصيص النموذج لاقتطاع جزء النص المطلوب.ويمكن توضيح المنهجية المستخدمة بالخطوات التالية [5]:

- تضمين الكلمات باستخدام BERT:

يتمثل دخل Bert بنص السؤال والمقاطع النصية الموافقة له، يجري جمع المقاطع النصية مع نص السؤال لتشكيل نص وحيد يمثل دخل النموذج، يجري تقطيع نص الدخل وفقاً لطريقة أجزاء الكلمات المشتركة shared workpieces [24]، كما يجري حذف التشكيل في النصوص العربية، ويعطي النموذج نتيجة تضمين كلمات النص (المتجهات الرقمية) كخرج نهائي للاعتماد عليه لاحقاً في مرحلة التخصيص.

- تخصيص نموذج BERT (Finetuning):

يتمثل دخل هذه المرحلة بنتيجة تضمين الكلمات الناتجة عن المرحلة السابقة، ويتمثل الخرج بالثنائية (i,j) والتي تعبر عن الجزء المقتطع المراد الحصول عليه بحيث تمثل i كلمة بداية الجزء النصي وتمثل j كلمة النهاية مع أخذ الشرط التالي بالاعتبار:

$$i \leq j \leq i + 15$$

---

[1] http://mazajak.inf.ed.ac.uk:8000/

[2] يعتبر **الكيان المسمى** كائنًا حقيقيًا مثل الأشخاص، والمواقع، والمؤسسات، والمنتجات وما إلى ذلك، والتي يمكن الإشارة إليه باسم علم. يمكن أن تكون هذه الكيانات مجردة أو لها وجود مادي. من الأمثلة على الكيانات المسماة باراك أوباما، مدينة نيويورك، فولكس فاجن جولف، أو أي شيء آخر يمكن تسميته. يمكن ببساطة النظر للكيانات المسماة على أنها مثال لكيان (على سبيل المثال، مدينة نيويورك هي مثال لمدينة).

جرت عملية التخصيص هذه عن طريق إضافة طبقات إضافية لشبكة BERT تأخذ كدخل نتيجة تضمين الكلمات ومهمتها إيجاد متجهين S ويمثل متجه كلمة البداية و E ويمثل متجه كلمة النهاية. تجري عملية التدريب بحيث يتم من أجل كل كلمة من كلمات النص i أخذ المتجه الناتج عن مرحلة التضمين T_i وحساب احتمالية كون i هي كلمة بداية وكذلك احتمال كونها كلمة نهاية كالتالي:

$$P_{start}(i) \propto \exp(S^T T_i)$$
$$P_{end}(i) \propto \exp(E^T T_i)$$

ويجري التدريب بحيث يتم تكبير قيمة العبارة التالية:

$$P_{start}(i) P_{end}(j)$$

وبذلك يتم الحصول على جزء قصير من المقطع النصي ويحمل الاحتمال الأكبر لورود الإجابة الصحيحة ضمنه. واعتماداً على المنهجية السابقة قمنا بوضع خوارزمية لاستخراج إجابات الأسئلة التعريفية كما يوضح الشكل (26).

الدخل: قائمة بثلاث مقاطع نصيّة (الناتجة عن مرحلة استخراج المقاطع النصيّة) S_p، ونص السؤال q.

الخرج: إجابة محددة باسم مكان، أو شخص، أو منظمة.

الخطوات:

1. تصنيف السؤال إلى أحد الأصناف التالية: سؤال عن مكان، أو سؤال عن شخص، أوسؤال عن منظمة ووضع النتيجة في q_{cls}.
2. إدخال نص السؤال q والمقاطع النصيّة S_p إلى نموذج BERT المخصص، واستخراج جزء من المقاطع النصية segment.
3. اذا كان صنف السؤال مكان q_{cls} == LOC:
   a. استخراج أول كيان مسمى من نوع مكان من النص segment وإعادته كإجابة.
4. وإلا اذا كان صنف السؤال شخص q_{cls} == PERS:
   a. استخراج أول كيان مسمى من نوع شخص من النص segment وإعادته كإجابة.
5. وإلا اذا كان صنف السؤال منظمة q_{cls} == ORG:
   a. استخراج أول كيان مسمى من نوع منظمة من النص segment وإعادته كإجابة.

الشكل 26 – خوارزمية الإجابة على الأسئلة التعريفية.

جرى استخدام نموذج Bert المستخدم ذاته في نظام SOQAL وذلك باستخدام الموارد العتادية المتاحة على منصة Google Collaboratory[1]، وقمنا باختيار هذا النموذج دون غيره نظراً لأن الرماز البرمجي المرتبط بعملية تدريبه وتخصيصه متاح ويمكن استخدامه والاستفادة منه، كما أن الزمن اللازم لتدريب النموذج قليل مقارنةً بغيره من النماذج مثل AraBert، على الرغم من

[1] يسمح Google Collaboratory لأي شخص بكتابة وتنفيذ كود Python من خلال المتصفح ، وهو مناسب بشكل خاص للتعلم الآلي وتحليل البيانات والتعليم.

أن نموذج AraBert متاح للاستخدام والتخصيص وأعطى نتائج أفضل من SOQAL إلا أنه يحتاج موارد عالية الأداء ولم نتمكن من تدريبه باستخدام الموارد المتاحة على منصة Google Collaboratory.

جرى استخدام أداة Arabic-NER[1] لتحقيق إمكانية التعرف على الكيانات المسمّاة في النصوص العربية، حيث تقوم هذه الأداة بالتعرف على ثلاثة أنواع من الكيانات وهي: المكان LOC، والأشخاص PERS، والمنظمات ORG. ويمكن لهذه الأداة التعرف على الكيانات التي تتألف من أكثر من كلمة كما في حال اسم "طلال عبد الهادي" حيث يجري التعرف عليها من قبل الأداة ويكون الخرج كالتالي "طلال: B-PERS، عبد:I-PERS، الهادي: I-PERS"، حيث B-PERS تعني أول كلمة من اسم الشخص وI-PERS تعني الكلمات اللاحقة من اسم الشخص. جرى تصنيف الأسئلة التعريفية لأحد الأصناف الثلاث (مكان – زمان – اشخاص) باستخدام مصنف آلات أشعة الدعم Support Vector Machine(SVM).

3.4.2.1.الإجابة عن أسئلة التعداد

تُعامل أسئلة التعداد معاملة الأسئلة التعريفية، بحيث تجري الإجابة على هذا النوع من الأسئلة باستخدام نفس الآلية المذكورة سابقاً مع اختلاف أنها تعيد كافة الكيانات الواردة ضمن الجزء النصي والموافقة لصنف السؤال المطروح عوضاً عن إعادة الكيان الأول فقط.

4.4.2.1.الإجابة عن أسئلة التلخيص

جرت الإجابة على هذا النمط من الأسئلة وفقاً للآلية ذاتها المطبقة في نظام SimBioNLQA والتي تقوم بإعادة المقاطع النصيّة الناتجة عن المرحلة السابقة (مرحلة استخراج المقاطع النصيّة)، بحيث تشكل المقاطع النصيّة ملخصاً لما تم السؤال عنه[7]، وقمنا بإضافة خطوة حذف المقاطع النصيّة المتكررة (التشابه بينها كبير جداً) وذلك لتحسين النتائج التي يعيدها النظام وجعلها أكثر منطقية، حيث جرى قياس التشابه بين المقاطع النصيَّة وفقاً لمقياس BM25، وعندما تكون قيمة التشابه بين مقطعين نصيّين أكبر أو تساوي 0.9 يتم حذف المقطع النصي الأصغر بينهما والاحتفاظ بالمقطع النصي الكبير (يقاس الطول بعدد الكلمات)، وذلك من منطلق أن المقطع الكبير قد يحوي معلومة غير واردة في الآخر؛ لذا تم اختياره ليبقى في الإجابة النهائية التي يعيدها النظام.

[1] https://github.com/HassanAzzam/Arabic-NER

2.الخاتمة

قمنا في هذا الفصل بعرض المنهجية المقترحة في بحثنا للإجابة عن الأسئلة المطروحة باللغة العربية الفصحى، حيث تم عرض المكونات الأساسية للنظام مع ذكر عمل كل مكوّن والخرج المتوقع منه والتقنيات والأدوات المستخدمة لتحقيقه، وسنعرض في الفصل التالي كيف جرى اختيار خطوات المعالجة المطبقة في كل مرحلة من مراحل النظام، وكذلك آليات اختبار كل مكوّن من مكونات النظام وآلية اختبار النظام بشكل عام والنتائج المرحلية والنهائية للمنهجية المقترحة.

الفصل الخامس: الاختبارات والنتائج

نقوم في هذا الفصل بشرح آلية اختبار المنهجية المقترحة، نبدأ بشرح عينات التدريب والاختبار المستخدمة، ومعايير التقويم المعتمدة، ونتائج اختبار الأنظمة الجزئية المكوّنة للنظام، ونقارن بالنهاية نتائج النظام بالنظم المشابهة له.

1.عينات التدريب والاختبار

نحتاج في بحثنا لعينتي تدريب، واحدة من أجل تدريب مصنف SVM في مرحلة تصنيف السؤال لأحد الأصناف الأربعة، والأخرى لتدريب نموذج BERT في مرحلة استخراج إجابات الأسئلة التعريفية. كما نحتاج لعينة اختبار من أجل اختبار كفاءة النظام والأنظمة الجزئية المكونة له.

1.1.عينة تدريب مصنف الأسئلة

جرى تدريب مصنف الأسئلة على عينة مكونة من 2300 سؤال منمّطة وفقاً للأنماط الأربعة المعتمدة في البحث وهي: أسئلة التأكيد، والأسئلة التعريفية، وأسئلة التعداد، وأسئلة التلخيص. بحيث جرى تجميع الأسئلة باستخدام google form (انظر الشكل(27)) وطرحه على مواقع التواصل الاجتماعي، وبذلك استطعنا الحصول على عينة التدريب المطلوبة والموّزعة كما يوضح الشكل(28).

طرح اسئلة باللغة العربية الفصحى

الهدف من ذلك الحصول على أكبر عدد ممكن من الاسئلة المطروحة "باللغة العربية الفصحى" بحيث يكون من أحد الأنواع المحددة (يمكنكم طرح أكثر من سؤال).

* Required

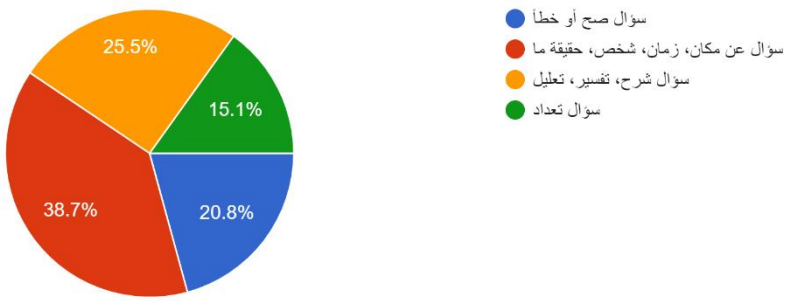
نوع السؤال الذي تريد طرحه *

- سؤال صح أو خطأ
- سؤال عن مكان، زمان، شخص، حقيقة ما
- سؤال شرح، تفسير، تعليل
- سؤال تعداد

السؤال الذي تريد طرحه *

Your answer

الشكل 27 – استبيان جمع عينة تدريب مصنف الأسئلة.



الشكل 28 – توزع عينة تدريب مصنف الأسئلة وفقاً للأنماط المعتمدة.

2.1.عينة تدريب نموذج BERT

جرى تدريب نموذج Bert المستخدم في مرحلة استخراج إجابات الأسئلة التعريفية باستخدام مجموعة بيانات Arabic Stanford Question Answering Dataset (Arabic SQuAD)[1]، والمؤلفة من 48344 سؤالاً مع المقاطع النصية المستخرجة منها والإجابات الموافقة لها، وهي ذاتها عينة التدريب المستخدمة في تدريب النماذج اللغوية في الأنظمة المشابهة.

3.1.عينة الاختبار

جرى اختبار نظامنا على مجموعة البيانات Arabic Reading Comprehension Dataset[2] (ARCD)، والمعتمدة ذاتها ضمن الدراسات المشابهة، وقمنا باختيارها كعينة اختبار للنظام نظراً لشموليتها من حيث اختلاف المجالات التي تتناولها، وتنوع أنماط الأسئلة المطروحة ضمنها، إضافة إلى أنها مجموعة البيانات المستخدمة في تقييم الأنظمة المشابهة (SOQAL - AraBert - AraELECTRA) ، وتتألف مجموعة البيانات هذه من 1395 ثنائية (سؤال، وجواب)، بحيث تم استخراج هذه الأسئلة 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)، ويوضح الشكل (29) بُنية مجموعة البينات المعتمدة كعينة اختبار للنظام. كما يوضح الجدول (6) توزّع الأسئلة المطروحة ضمن المجالات المختلفة في عينة الاختبار.

---

[1] https://github.com/husseinmozannar/SOQAL

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

```
file.json
├── "data"
│   └── [i]
│       ├── "paragraphs"
│       │   └── [j]
│       │       ├── "context": "paragraph text"
│       │       └── "qas"
│       │           └── [k]
│       │               ├── "answers"
│       │               │   └── [l]
│       │               │       ├── "answer_start": N
│       │               │       └── "text": "answer"
│       │               ├── "id": "<uuid>"
│       │               └── "question": "paragraph question?"
│       └── "title": "document id"
└── "version": 1.1
```

الشكل 29 – البنية العامة لعينة اختبار النظام.

عدد الأسئلة الكلي	عدد المستندات النصية الكلي	عدد المجالات	عدد المستندات في المجال الواحد	عدد المقاطع النصية في المستند الواحد	عدد الأسئلة المُستخرجة من المستند الواحد	عدد الأسئلة المُستخرجة من المقطع النصي الواحد
1395	465	155	3	3	3	1

الجدول 6 – توزع الأنماط المختلفة للأسئلة ضمن عينة الاختبار.

2.معايير التقويم

جرى تقييم النظام وفقاً لقياس توافق الجملة Sentence Match (SM) و قياس Macro F1 Score وذلك من أجل أسئلة التعداد والأسئلة التعريفية، وجرى اختيار هذه القياسات نظراً لأنها القياسات المعتمدة في الأبحاث المشابهة والمعتمدة على نفس عينة الاختبار، وبذلك نستطيع مقارنة نظامنا بهذه النظم. يجري حساب SM وفقاً للمعادلة التالية [5]:

$$SM = \frac{nb\ of\ matched\ answers}{nb\ of\ all\ answers}$$

والذي يمثل نسبة الإجابات التي ترد في نفس جملة الجواب الصحيح (المرفقة في عينة الاختبار) من مجمل الإجابات.

ويجري حساب Macro F1 Score بحساب متوسط الكلمات المتطابقة من الإجابات التي يعيدها النظام مع الإجابات الصحيحة (المرفقة في عينة الاختبار) [5].

3.نتائج اختبار النظام

من أجل كل ثنائية (سؤال، وجواب) في عينة الاختبار المعتمدة ARCD، قمنا بطرح السؤال على النظام المقترح، ومقارنة إجابة النظام بالإجابة الصحيحة المرفقة في عينة الاختبار، وكانت نتائج النظام كما يوضح الجدول(7).

F1	SM
62.5%	92.4%

الجدول 7 – نتائج اختبار النظام.

بمقارنة النظام المقترح بالنظم المشابهة والمعتمدة على نفس عينة الاختبار، كانت النتائج كما يوضح الجدول(8).

معيار المقارنة/النظام		SOQAL	AraBert	AraELECTRA	النظام المقترح
لغة المعالجة		العربية الفصحى	العربية الفصحى	العربية الفصحى	العربية الفصحى
مجال المعطيات		مفتوحة المجال	مفتوحة المجال	مفتوحة المجال	مفتوحة المجال
أنماط الأسئلة المعالجة		غير محددة النمط	غير محددة النمط	غير محددة النمط	غير محددة النمط
شكل الخرج		مقطع نصي	مقطع نصي	مقطع نصي	**جواب محدد وفقاً لنمط السؤال المطروح**
آلية الاختبار والنتائج	مجموعة البيانات	ARCD	ARCD	ARCD	ARCD
	SM	90%	92%	-	**92.4%**
	F1	61.3%	62.7%	**71.22%**	62.5%

الجدول 8 – مقارنة النظام المقترح بنظام SOQAL.

نلاحظ من النتائج السابقة أن النظام المقترح استطاع التفوق على الأنظمة المشابهة وفقاً لقياس SM، كما أن هذه الأنظمة تفتقر لوجود مرحلة معالجة نمط السؤال والإجابة وفقاً له، فهي تقتصر على اقتطاع جزء من النص يحتوي على الإجابة الصحيحة وتعيده كإجابة نهائية، **وهذا ما تم تداركه في نظامنا**، حيث يستطيع النظام المقترح تقديم إجابة واضحة ومحددة وفقاً لنمط السؤال المطروح. كما استطاع نظامنا تحقيق نسبة 62.5% من أجل قياس F score وبالمقارنة مع النظم المشابهة نلاحظ أن نظامنا لا يتفوق عليها ويعود السبب في ذلك إلى كون قياس Fscore يأخذ متوسط الكلمات المتطابقة مع كلمات الإجابات المرفقة في عينة الاختبار، ولكن الإجابات في عينة الاختبار المعتمدة هي عبارة عن مقاطع نصية بينما يعيد نظامنا إجابات محددة (على خلاف الأنظمة السابقة) وبالتالي عدد الكلمات المتطابقة قليل (تبعا لإجابات نظمنا المختصرة). يوضح الجدول(9) بعض الأمثلة عن إجابات نظام SOQAL (والتي تشبه في صيغتها نظام AraBert و AraELECTRA) ومقارنة هذه الإجابات بإجابات نظامنا المقترح والتي تبين قدرة نظامنا في تقديم إجابة محددة وأكثر دقة وفقاً لنوع السؤال.

السؤال	نوع السؤال	إجابة نظام SOQAL	إجابة نظامنا
هل ارتفاع السكري عند الحوامل أمر شائع؟	سؤال تأكيد	ازداد إلى أكثر من الضعف	نعم
ماهي الدولة التي تحد شمال الولايات المتحدة؟	سؤال تعريفي	كندا شمالاً والمكسيك جنوباً	كندا
من هو جمال أحمد حمزة خاشقجي؟	سؤال تلخيص	صحفي وإعلامي سعودي	جمال أحمد حمزة خاشقجي (13 أكتوبر 1958، المدينة المنورة – 2 أكتوبر 2018)، صحفي وإعلامي سعودي، رأس عدّة مناصب لعدد من الصحف في السعودية، وتقلّد منصب مستشار، كما أنّه مدير عام قناة العرب الإخبارية سابقًا

الجدول 9 – مقارنة إجابات النظام المقترح بإجابات نظام SOQAL

نوضح في الملحق (1) بعض الأمثلة من عينة الاختبار، ونبيّن النتائج المرحلية والنهائية للنظام المقترح.

4.اختبار الأنظمة الجزئية

قمنا باختبار كل مرحلة من مراحل النظام بشكل منفصل، وذلك من أجل معرفة كفاءة المنهجية المستخدمة في إنجاز كل مرحلة بقطع النظر عن ارتباطها بكفاءة المراحل السابقة لها. الغاية الأساسية من هذه الفقرة هي تعليل اختيار خطوات المعالجة المطبقة في كل مرحلة من مراحل النظام وذلك وفقاً للتجريب وقياس الدقة. لم نتمكن من مقارنة النتائج المرحلية لنظامنا بالدراسات المشابهة وذلك لعدم وجود نتائج مرحلية في الأعمال المشابهة الموجودة في دراستنا المرجعية.

1.4.اختبار مرحلة تحليل السؤال

يجري في مرحلة تحليل السؤال معالجة السؤال لغوياً، وتصنيفه لأحد الأصناف الأربعة، وتوسيع السؤال بإضافة مرادفات لكلمات السؤال. جرى إنجاز كل من هذه الخطوات وفقاً للتجريب وقياس الدقة عند كل خطوة، حيث جرى قياس دقة هذه المرحلة وفقاً لنسبة الأسئلة التي تم تصنيفها بشكل صحيح من مجمل الأسئلة الواردة في مجموعة البيانات.

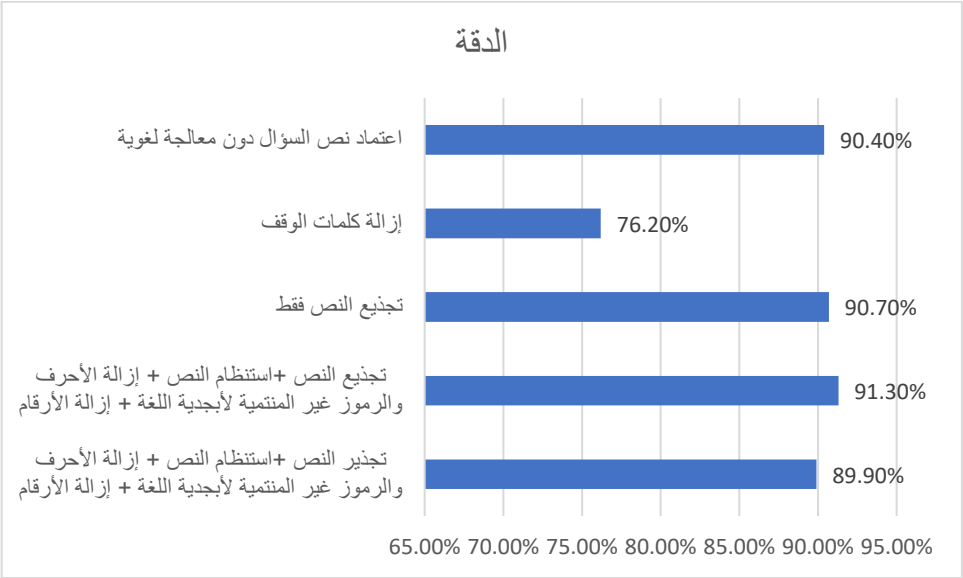
1.1.4.معالجة السؤال لغوياً

جرى تحديد خطوات معالجة السؤال لغوياً بتجريب عدة خطوات وهي:

- تجذير كلمات السؤال.
- تجذيع كلمات السؤال.

- استنظام نص السؤال.
- إزالة الأحرف والرموز غير المنتمية لأبجدية اللغة.
- إزالة الأرقام.
- إزالة كلمات الوقف.

قمنا بقياس الدقة عند كل خطوة، واخترنا الخطوات التي تعطي أكبر دقة تصنيف كما يوضح الشكل (30).



الشكل 30- قياس دقة مرحلة تحليل السؤال وفقاً لخطوات المعالجة اللغوية المعتمدة.

وفقاً للنتائج السابقة، تبين أن النظام تقل كفاءته عند حذف كلمات الوقف؛ ويعود ذلك لأن أغلب أدوات الاستفهام – والتي تمييز نوع السؤال – تعامل معاملة كلمات الوقف ويتم حذفها. وبالتالي لن يستطيع المصنف التعرف على النمط الصحيح للسؤال. كما أن استخدام جذر الكلمة بدلاً من جذعها قلل من كفاءة النظام وذلك لأن استخدام الجذر يوحد الكثير من الكلمات التي يمكن لها أن تمييز السؤال.

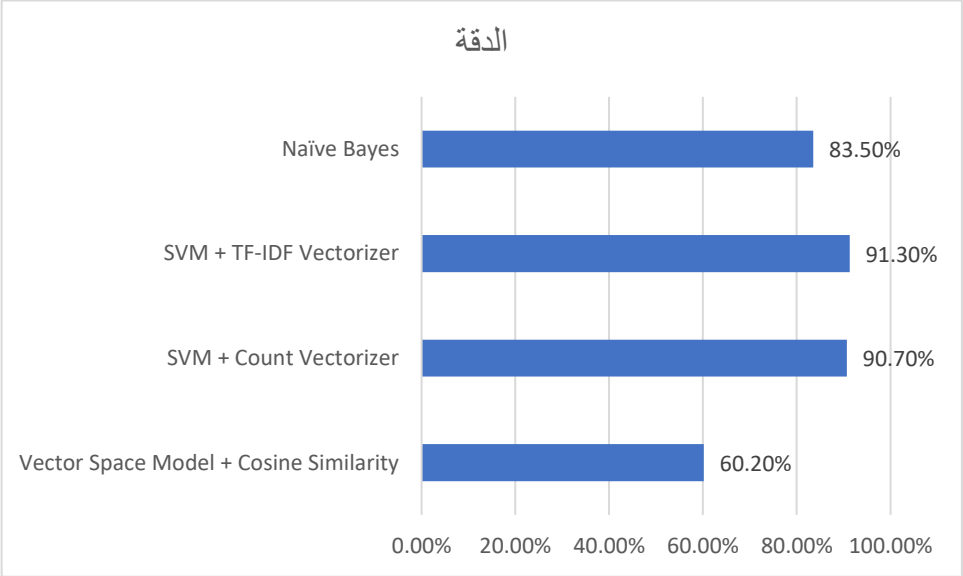
2.1.4.تصنيف السؤال

قمنا باختبار عدة مصنفات لتحقيق مرحلة تصنيف السؤال وهي:

- التصنيف باستخدام نموذج فضاء المتجهات Vector Space Model(VSM) مع قياس جيب التمام بين المتجهات.
- التصنيف باستخدام مصنف Support Vector Machine (SVM) مع استخدام شعاع التكرار كشعاع ميزات Features Vector، وقمنا باختيار هذا المصنف نظراً لأنه المصنف المستخدم في الأنظمة المشابهة وأثبت كفاءته في هذه المرحلة (مثل نظام SimBioNLQA).

- التصنيف باستخدام مصنف Support Vector Machine (SVM) مع استخدام شعاع قياس TF-IDF كشعاع ميزات.
- التصنيف باستخدام مصنف Naïve Bayes مع استخدام شعاع قياس TF-IDF كشعاع ميزات.

قمنا بقياس دقة التصنيف عند كل مصنف من المصنفات السابقة، وكانت النتائج كما يوضح الشكل(31).



الشكل 31 – قياس دقة النظام وفقاً لنوع مصنف السؤال.

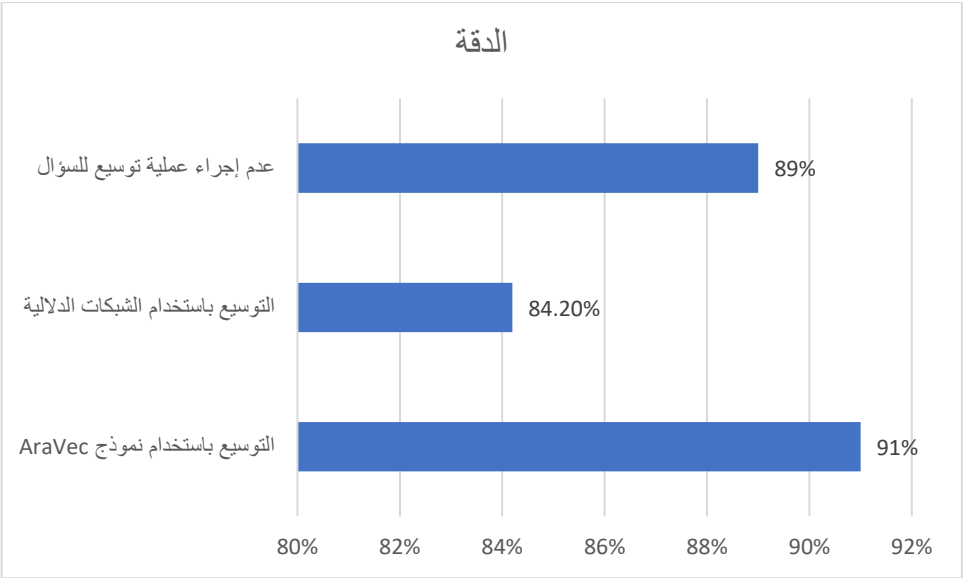
وفقاً للنتائج السابقة، تبين أن النظام يعطي أفضل دقة باستخدام مصنف SVM مع شعاع TF-IDF كشعاع ميزات، لذا تم اعتماده في مرحلة تصنيف السؤال.

3.1.4.توسيع السؤال

يتم توسيع السؤال بإضافة مرادفات لكلماته، قمنا بتجريب آليتين مختلفتين للقيام بذلك وهما:

- التوسيع باستخدام الشبكات الدلالية.
- التوسيع باستخدام نموذج تضمين الكلمات AraVec.

قمنا بقياس الدقة عند كل آلية، وكانت النتائج كما يوضح الشكل(32).



الشكل 32 – قياس دقة النظام وفقاً لآلية توسيع السؤال المعتمدة.

وفقاً للنتائج السابقة، تبين أن النظام يعطي أفضل دقة عند توسيع السؤال باستخدام نموذج تضمين الكلمات AraVec 1.0، ويعود ذلك لكون نماذج تضمين الكلمات هذه تعيد المرادفات التي ترد عادة بنفس السياق وليس المرادفات المعجمية فقط كما في حال استخدام الشبكات الدلالية.

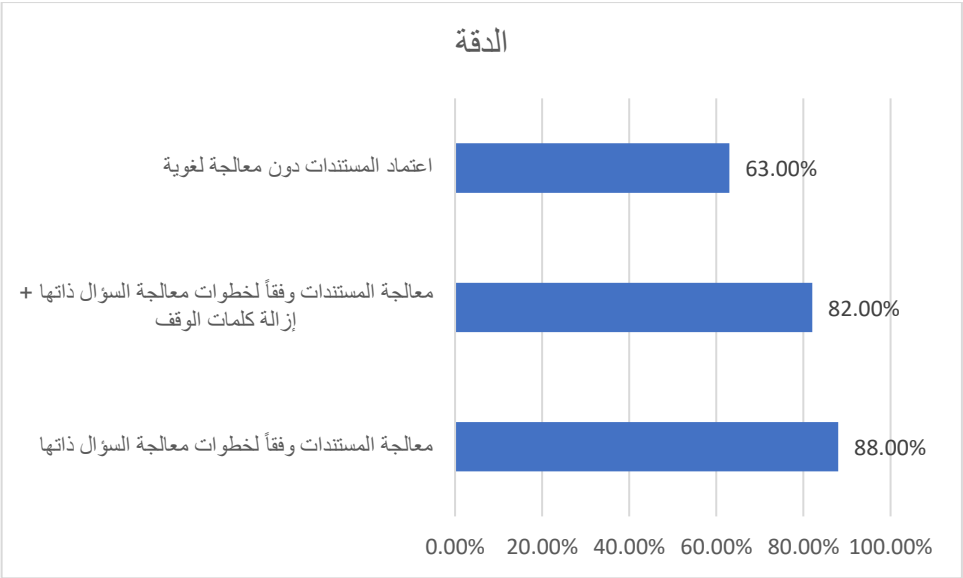
جرى اعتماد AraVec في تحقيق مراحل تضمين الكلمات في هذا البحث وذلك لأن هذه الأداة تحتوي على ستة نماذج تضمين كلمات مختلفة وتم تحقيقها استناداً لمصادر متنوعة وهي:تويتر، وويكيبديا، وصفحات الوب. ويصل عدد الكلمات التي يمكن تضمينها باستخدام نماذج AraVec لأكثر من 3,300,000,000 كلمة. وجرى اختيار نموذج AraVec المعتمد على مقالات ويكيبديا وذلك لأنها مبنية على بيانات باللغة العربية الفصحى ومتنوعة من ناحية المجال الذي تتناوله وهذا ما يتناسب مع نظامنا [12].

2.4.اختبار مرحلة استخراج المستندات

تجري في مرحلة استخراج المستندات معالجة لغوية على هذه المستندات، حيث جرى تحديد خطوات معالجة المستندات هذه بتجريب عدة خطوات وقياس الدقة عند كل خطوة:

- معالجة المستندات وفقاً لخطوات معالجة السؤال ذاتها.
- معالجة المستندات وفقاً لخطوات معالجة السؤال ذاتها إضافةً لحذف كلمات الوقف.
- عدم إجراء أي معالجة لغوية على المستندات واعتمادها كما هي في مرحلة استخراج المستندات.

قمنا بقياس دقة النظام عند كل خطوة، حيث جرى قياس دقة هذه المرحلة بنسبة المستندات المسترجعة بشكل صحيح من العدد الكلي للمستندات المسترجعة، وكانت النتائج كما يوضح الشكل(33).



الشكل 33 – قياس دقة النظام وفقاً للمعالجة اللغوية المعتمدة في مرحلة استخراج المستندات.

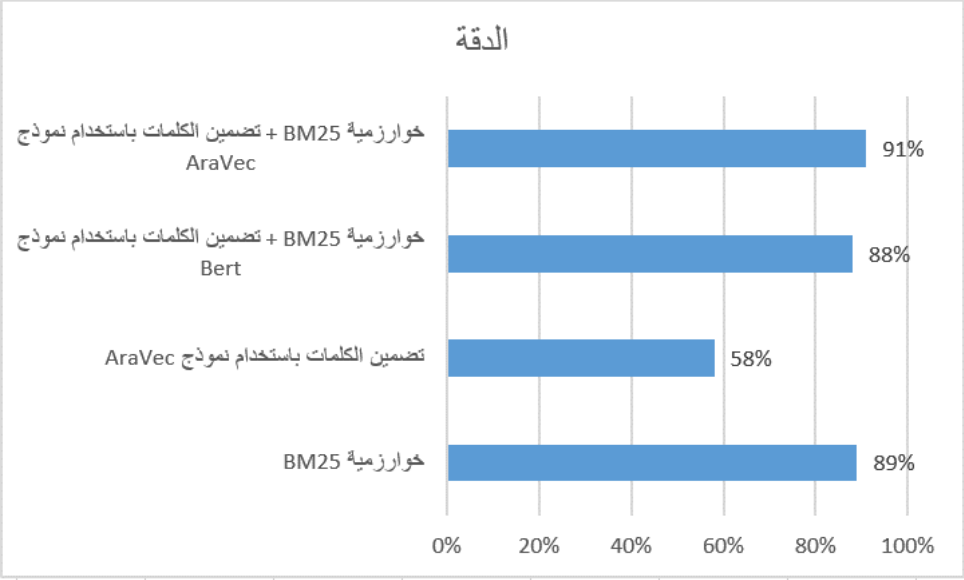
وفقاً للنتائج السابقة، تبين أن مرحلة إزالة كلمات الوقف من المستندات قد زادت من دقة النظام بشكل كبير؛ لذلك جرى اعتماد هذه الخطوة في معالجة المستندات. وجرى استرجاع أول خمسة مستندات مرتبطة بالسؤال كدخل للمرحلة اللاحقة.

3.4.اختبار مرحلة استخراج المقاطع النصيّة

قمنا بتجريب عدة آليات لإنجاز مرحلة استخراج المقاطع النصيّة وهي:

- استخراج المقاطع النصية باعتماد خوارزمية BM25.
- استخراج المقاطع النصية بتطبيق خوارزمية معتمدة على نموذج تضمين الكلمات AraVec.
- استخراج المقاطع النصية بتطبيق خوارزمية معتمدة على نموذج تضمين الكلمات Bert.
- استخراج المقاطع النصية باعتماد نموذج خطي يجمع بين خوارزمية BM25 ونموذج AraVec.

قمنا بقياس دقة النظام عند كل آلية، حيث جرى قياس دقة هذه المرحلة بنسبة المقاطع النصية المسترجعة بشكل صحيح من العدد الكلي للمقاطع النصيّة المسترجعة، وكانت النتائج كما يوضح الشكل(34).



الشكل 34 – قياس دقة النظام وفقاً للآلية المتبعة في استخراج المقاطع النصيّة.

وفقاً للنتائج السابقة، تبين أن النظام يعطي أفضل دقة عند استخراج المقاطع النصيّة وفقاً للآلية التي تعتمد على الجمع ما بين خوارزمية BM25 ونموذج تضمين الكلمات AraVec في نموذج خطي واحد؛ لذلك جرى اعتماد هذا النموذج في تحقيق مرحلة استخراج المقاطع النصيّة.

4.4.اختبار مرحلة استخراج الإجابة

جرى اختبار مرحلة استخراج الإجابة بقياس دقة كل خوارزمية من الخوارزميات الأربعة المطبقة في هذه المرحلة وهي: خوارزمية استخراج إجابات أسئلة التأكيد، وخوارزمية استخراج إجابات الأسئلة التعريفية، وخوارزمية استخراج إجابات أسئلة التعداد، وخوارزمية استخراج إجابات أسئلة التلخيص، حيث جرى قياس دقة هذه المرحلة بنسبة الأسئلة المجاب عنها بشكل صحيح من مجمل الأسئلة المطروحة، ويعتبر الجواب صحيحاً وفقاً للمراجعة البشرية للأجوبة التي يقدمها النظام ومقارنتها بالجواب الصحيح المرفق في عينة الاختبار. يوضح الجدول(10) دقة كل خوارزمية من الخوارزميات السابقة في استخراج الإجابات الصحيحة وفقاً لنمط السؤال المطروح.

الخوارزمية	عدد الأسئلة المطروحة	عدد الأسئلة المجاب عنها بشكل صحيح	الدقة
خوارزمية استخراج إجابات أسئلة التأكيد	36	26	72%
خوارزمية استخراج إجابات الأسئلة التعريفية	171	124	72.5%
خوارزمية استخراج إجابات أسئلة التعداد	25	16	64%
خوارزمية استخراج إجابات أسئلة التلخيص	1163	1055	90.7%

الجدول 10 – دقة خوارزميات استخراج الإجابات.

5.تحليل النتائج

بيّنت الاختبارات السابقة أن النظام المقترح استطاع تحقيق نتائج جيدة في الإجابة عن الأنماط الأربعة من الأسئلة (أسئلة التأكيد، وأسئلة التلخيص، وأسئلة التعداد، والأسئلة التعريفية)، وبمقارنة النظام مع النظم المشابهة السابقة التي اهتمت باللغة العربية بشكل خاص، نلاحظ أن نظامنا المقترح استطاع التفوق عليها في واحدة من النقاط التالية على الأقل:

- قدرة نظامنا على الإجابة على أنماط مختلفة من الأسئلة.
- قدرة نظامنا على الإجابة عن أسئلة مطروحة ضمن مجالات مختلفة.
- قدرة نظامنا على الإجابة عن الأسئلة المطروحة بشكل دقيق وفقاً لنمط السؤال المطروح.

فمن خلال دراستنا المرجعية نلاحظ أنه بالرغم من النتائج الجيدة التي حققتها نظم الإجابة الآلية باللغة العربية؛ إلا أنها كانت محدودة إما بالمجال، أو بنمط السؤال، أو بكيفية الإجابة. استطاع نظامنا المقترح أن يحقق نتائج جيدة مع أخذ النقاط السابقة بالاعتبار، والسبب في ذلك هو المنهجيات المختلفة التي قمنا باتباعها على مستوى كل مرحلة من مراحل النظام، فالسبب في قدرة النظام على الإجابة عن الأسئلة المطروحة ضمن مجالات مختلفة يعود للمنهجية المستخدمة في قياس تشابه النصوص في مرحلة استخراج المقاطع النصيّة والتي تأخذ بالاعتبار قياس التشابه وفقاً للدلالة مما يزيد من دقة هذه المرحلة. إن الخوارزميات المقترحة والتي جرى تحقيقها في مرحلة استخراج الإجابة هي السبب في قدرة النظام على الإجابة عن الأسئلة المطروحة بشكل دقيق وفقاً لنمط السؤال المطروح، حيث قمنا باعتماد منهجيات وتقنيات مختلفة في معالجة كل نمط بشكل منفصل مما ساهم في الحصول على إجابة دقيقة وفقاً لهذا النمط، مع العلم أن هذه المنهجية غير مستخدمة بعد في الأنظمة المشابهة العربية.

عند مقارنة نظامنا المقترح بنظام SOQAL (والذي يشبه في شكل الإجابة التي يعيدها نظامي AraBert و AraELECTRA) نلاحظ أن كلا النظامين يجيبان عن الأسئلة المطروحة باللغة العربية وبقطع النظر عن مجال السؤال وعن نمطه وباعتماد عينة الاختبار ذاتها في كلا البحثين وهي ACRD، نلاحظ أن نظامنا المقترح يتفوق على SOQAL في قدرته على تقديم إجابة دقيقة، كما في مثال طرح سؤال التأكيد "هل ارتفاع السكري عند الحوامل أمر شائع؟"، يجيب نظام SOQAL بالمقطع النصي "ازداد إلى أكثر من الضعف" والذي يوحي للقارئ بالإجابة الصحيحة، بينما يستطيع نظامنا من خلال مرحلة تحليل المشاعر في المقاطع النصية المستخرجة ونص السؤال من تقديم إجابة أكثر دقة وهي "نعم"، وكذلك الحال بالنسبة للأسئلة التعريفية وأسئلة التعداد، فعند طرح السؤال "ماهي الدولة التي تحد شمال الولايات المتحدة؟" على كلا النظامين، يجيب نظام SOQAL بالمقطع النصي "كندا شمالاً والمكسيك جنوباً"، بينما استطاع نظامنا ومن خلال إضافة مرحلة التعرف على الكيانات المسماة من الإجابة بشكل أدق باسم الدولة فقط وهو "كندا"، وفي حال أسئلة التلخيص مثل سؤال "من هو جمال أحمد خاشقجي؟" يجيب نظام SOQAL بإجابة مختصرة وهي "صحفي وإعلامي سعودي" بينما يستطيع نظامنا تقديم ملخص بسيط عن الشخص الذي تم السؤال عنه (انظر الإجابة في الجدول (9)) مما يتناسب بشكل أكبر مع هذا النمط من الأسئلة.

بشكل عام، استطاع النظام المقترح الإجابة بشكل صحيح على الأسئلة المطروحة على اختلاف أنماطها وجرى تقييم النظام وفق معياري Sentence Match(SM) و Macro F-Measure واستطاع النظام تحقيق النتائج SM = 92.4% و F1 = 62.5%، وبتفصيل نتائج النظام وفقاً لنمط السؤال ، (انظر الجدول (8))، نلاحظ أن النظام قد حقق أعلى دقة في الإجابة عن أسئلة التلخيص والتي وصلت لـ 90.7%، ويعود السبب في ذلك إلى أن النظام يجيب على هذا النمط من الأسئلة باعتماد المقاطع النصيّة الناتجة عن مرحلة استخراج المقاطع النصيّة كملخص عن الإجابة، لذلك، تقاس دقة الإجابة عن هذا النمط من الأسئلة بدقة مرحلة استخراج المقاطع النصيّة. أما في حال أسئلة التأكيد فقد استطاع النظام الإجابة على 26 سؤالاً بشكل صحيح من أصل 36 سؤال، وفشل في الإجابة عن 10 منها، والسبب في ذلك يعود لأداة تحليل المشاعر المستخدمة ومجموعة البيانات المدربة عليها، حيث تم الاعتماد في نظامنا على أداة Mazajk والتي تم تدريبها على مجموعة بيانات مجمّعة من تغريدات توتير وباللغة العامية، وبالتالي تقوم هذه الأداة بالتصنيف تبعاً للمشاعر العامة كما في حال كلمة "مرض" أو "سرقة" فهي تقوم بتصنيفهم على أنهم كلمات سلبية، بينما نحتاج في نظامنا لمصنف يعتمد فقط على النفي في الجمل مثل ورود كلمة "لا" أو "لم" لاعتبارها جملة تحمل مشاعر سلبية. وبمراجعة نتائج النظام عند الأسئلة التعريفية وأسئلة التعداد، نلاحظ أن النظام استطاع أن يجيب بشكل صحيح على 140 سؤال من أصل 191 سؤال، وفشل في الإجابة عن 51 سؤالاً، يمكن في هذه الحالة أن يعود السبب إما لمرحلة اقتطاع الجزء النصي باستخدام شبكات Bert، أو لمرحلة تصنيف نوع الكيان المسمى لأحد الأصناف الثلاثة (مكان، أو شخص ، أو منظمة)، مثال في السؤال "من هو سيد الرجال؟"، تعيد مرحلة اقتطاع الجزء النصي "رسول الله" وعند تمرير المقطع النصي لمرحلة استخراج الكيان المسمى؛ لن يتم التعرف على أي كيان مسمى ضمن الجملة السابقة وسيعيد النظام جملة فارغة ""، بينما الإجابة الصحيحة المتوقعة هي "أبو القاسم محمد بن عبد الله"، وكذلك الأمر في حال التصنيف الخاطئ لنوع الكيان، كما في السؤال "أين عاش جمال خاشقجي؟" والذي تم تصنيفه على أنه سؤال عن شخص، وبالتالي تم استخراج الكيان المسمى من نوع شخص بينما السؤال في الحقيقة هو عن مكان، لذلك، يمكن التحسين من دقة الإجابة على هذه الأنماط من الأسئلة بتحسين آلية اقتطاع الجزء النصي وكذلك المصنف المستخدم في مرحلة تصنيف الكيان المسمى المراد استخراجه كإجابة، ويمكن تحسين آلية اقتطاع الجزء النصي ببناء نموذج لغوي شبيه بالنماذج اللغوية المستخدمة لهذا الغرض مثل (AraBert) وتدريبها على مجموعات واسعة من البيانات بحيث نستطيع تحقيق دقة أفضل من تلك المحققة في الأنظمة المشابهة، ويمكن تحسين آلية تصنيف الكيانات المسماة ببناء مصنف وتدريبه على بيانات متخصصة في نظام الإجابة الآلية بدلاً من استخدام أداة عامة لهذا الغرض.

الفصل السادس: الخاتمة والآفاق المستقبلية

انطلاقاً من الغاية الأساسية للبحث، وهي الحصول على نظام إجابة آلية باللغة العربية مفتوح المجال وغير محدد بأنماط أسئلة معينة وقادر على توفير إجابات دقيقة وواضحة وفقاً لنمط السؤال المطروح؛ قدمنا منهجية جديدة لتحقيق نظام إجابة آلية يوفر هذه الغاية قدر الإمكان، بحيث اعتمدت المنهجية المقترحة على أربعة مراحل رئيسية وهي: مرحلة معالجة السؤال، ومرحلة استخراج المستندات، ومرحلة استخراج المقاطع النصية، ومرحلة استخراج الإجابة. وارتكز البحث على إيجاد خوارزميات جديدة وفعّالة لتحقيق مرحلتي استخراج المقاطع النصيّة واستخراج الإجابة، ويعود ذلك لتأثيرهما الكبير على فعالية النظام وكفاءته، حيث جرى تحقيق هاتين المرحلتين باستخدام تقنيات معالجة لغوية حديثة، حيث تم الاستفادة من النماذج اللغوية المستخدمة مؤخراً في أنظمة الإجابة الآلية وذلك للتحسين من نتائج مرحلة استخراج المقاطع النصيّة، كما جرى تحقيق مرحلة استخراج الإجابة باستخدام تقنية التعرف على الكيانات المسماة وذلك للإجابة عن الأسئلة من نمط الأسئلة التعريفية وأسئلة التعداد، واعتماد تقنية تحليل المشاعر للإجابة عن أسئلة التأكيد، وفي أسئلة التلخيص جرى اعتماد خرج مرحلة استخراج المقاطع النصيّة مع حذف التكرار، حيث اعتمدنا بشكل عام في مرحلة استخراج الإجابة على بناء خوارزميات مستوحاة من تلك المستخدمة مؤخراً في الأنظمة المشابهة باللغة الإنكليزية والتي أثبتت كفاءتها في ذلك.

بيّنت الاختبارات التي أجريناها أن المنهجية المقترحة استطاعت الإجابة بشكل صحيح على أسئلة مطروحة باللغة العربية الفصحى ضمن مجالات مختلفة وبأنماط عدة، حيث جرى تقييم النظام وفقاً لمعياري SM و F1 واستطاع النظام تحقيق النتائج SM = 92.4% و F1 = 62.5%، كما استطاعت المنهجية المقترحة من التفوق على النظم المشابهة وفقاً لقياس SM وأيضاً في قدرتها على استخراج إجابات دقيقة وفقاً لنمط السؤال المطروح.

تدفعنا النتائج الجيدة التي حصلنا عليها في هذا البحث من متابعة العمل ضمن نفس المجال، وذلك من خلال إجراء عدة تحسينات، ويمكن تلخيصها في النقاط التالية:

- **تحسين آلية تحليل المشاعر المستخدمة في مرحلة استخراج إجابات التأكيد:**

لاحظنا من خلال اختبار النظام وتحليل نتائجه مدى تأثير مرحلة تحليل المشاعر على دقة النظام في استخراج إجابات التأكيد، وبالتالي يمكن العمل على هذه المرحلة من خلال تحقيق منهجية جديدة في تحليل المشاعر مخصصة لحالة نظامنا

(نظام إجابة آلية) بدلاً من اعتماد الأدوات العامة والتي قد تكون مخصصة لأغراض أخرى (كما في حالة تحليل مشاعر تغريدات تويتر).

- **توسيع نطاق الأسئلة التعريفية:**

يعالج نظامنا الأسئلة التعريفية التي يُستفسر بها عن مكان، أو شخص، أو منظمة فقط. وبالتالي يمكن توسيع نطاق الأسئلة التعريفية (وكذلك أسئلة التعداد) لتشمل الأسئلة عن الزمان، والأحداث، والتواريخ، والأرقام وغيرها. ويمكن تحقيق ذلك من خلال تطوير أدوات التعرف على الكيانات المسماة المستخدمة في هذه المرحلة لتشمل الأنماط الإضافية المطلوبة.

- **تحسين آلية استخراج إجابات أسئلة التلخيص:**

تجري آلية الإجابة عن أسئلة التلخيص، في نظامنا، باعتماد خرج مرحلة استخراج المقاطع النصيّة كما هي مع حذف المكرر منها. ويمكننا التحسين من جودة الملخص الذي يشكل الإجابة من خلال استخدام تقنيات التلخيص الآلي.

المراجع

[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, no. 3 (2012).

[2] Azmi, Aqil M., and Nouf A. Alshenaifi. "Lemaza: An Arabic why-question answering system." Natural Language Engineering 23, no. 6 (2017): 877.

[3] Abdi, A., Hasan, S., Arshi, M., Shamsuddin, S. M., & Idris, N. (2020). A question answering system in hadith using linguistic knowledge. Computer Speech & Language, 60, 101023.

[4] Abu Taha, Alaa W. "An ontology-based Arabic question answering system." (2015).

[5] Mozannar, Hussein, Karl El Hajal, Elie Maamary, and Hazem Hajj. "Neural arabic question answering." arXiv preprint arXiv:1906.05394 (2019).

[6] Al-Smadi, Mohammad, Islam Al-Dalabih, Yaser Jararweh, and Patrick Juola. "Leveraging Linked Open Data to Automatically Answer Arabic Questions." IEEE Access 7 (2019): 177122-177136.

[7] 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.

[8] Jurafsky, Dan. Speech & language processing. Pearson Education India, 2000.

[9] Gomaa, Wael H., and Aly A. Fahmy. "A survey of text similarity approaches." International Journal of Computer Applications 68, no. 13 (2013): 13-18.

[10] Li, Yang, and Tao Yang. "Word embedding for understanding natural language: a survey." In Guide to big data applications, pp. 83-104. Springer, Cham, 2018.

[11] Gupta, L. (2021, January 7). Differences Between Word2Vec and BERT - The Startup. Medium. https://medium.com/swlh/differences-between-word2vec-and-bert-c08a3326b5d1

[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] Horev, Rani. "BERT Explained: State of the art language model for NLP." Towards Data Science, Nov 10 (2018).

[14] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[15] Antoun, Wissam, Fady Baly, and Hazem Hajj. "Arabert: Transformer-based model for arabic language understanding." arXiv preprint arXiv:2003.00104 (2020).

[16] Chen, Danqi, Adam Fisch, Jason Weston, and Antoine Bordes. "Reading wikipedia to answer open-domain questions." *arXiv preprint arXiv:1704.00051* (2017).

[17] Chaybouti, Sofian. "EfficientQA: a RoBERTa Based Phrase-Indexed Question-Answering System." *arXiv preprint arXiv:2101.02157* (2021).

[18] Al-Khawaldeh, Fatima T. "Answer extraction for why Arabic questions answering systems: EWAQ." *arXiv preprint arXiv:1907.04149* (2019).

[19] Ahmed, Waheeb, and P. Babu Anto. "A Hybrid Question Answering System." *Current Journal of Applied Science and Technology* (2019): 1-7.

[20] Antoun, Wissam, Fady Baly, and Hazem Hajj. "AraELECTRA: Pre-Training Text Discriminators for Arabic Language Understanding." *arXiv preprint arXiv:2012.15516* (2020).

[21] What is TF-IDF? (2019, May 10). MonkeyLearn Blog. https://monkeylearn.com/blog/what-is-tf-idf

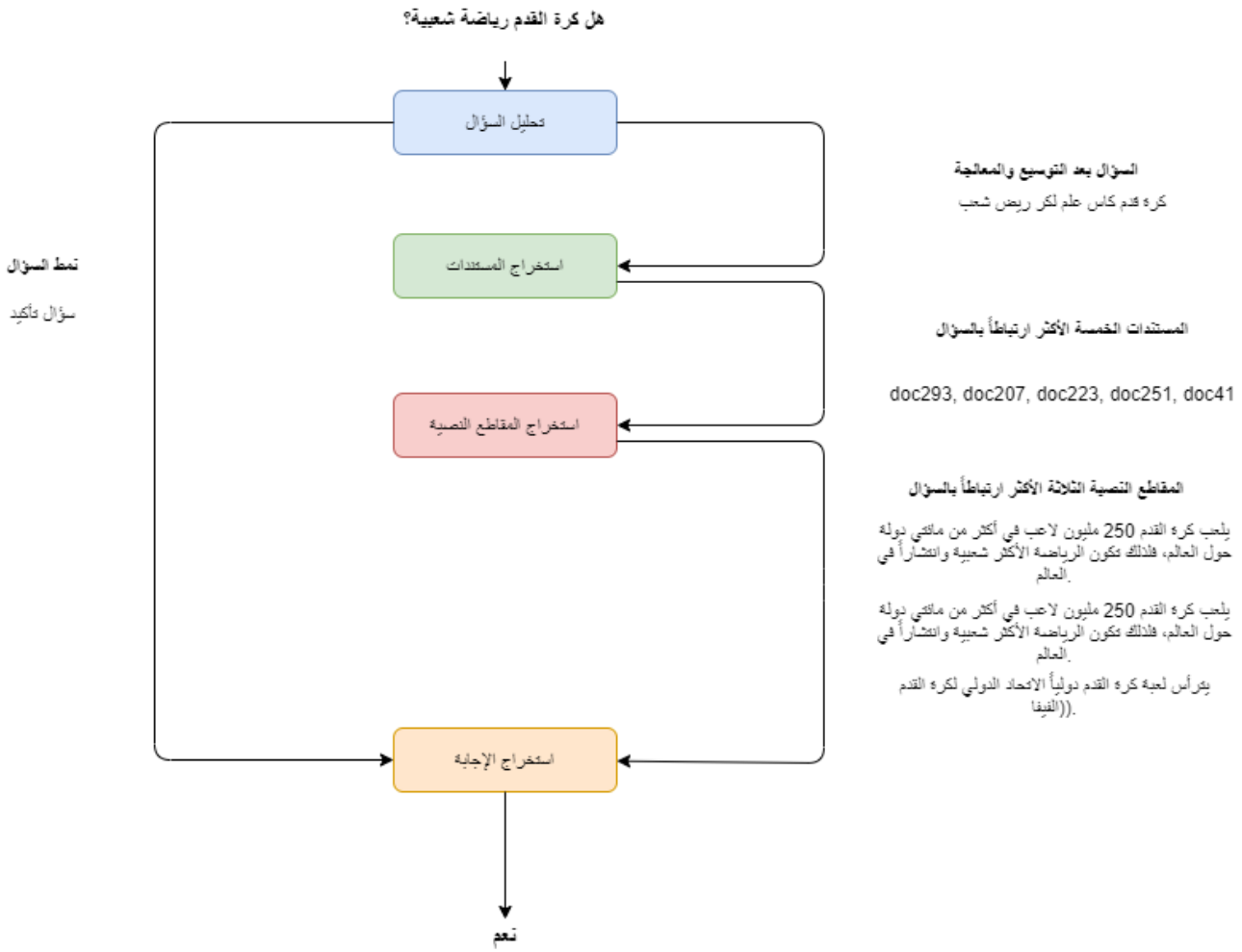
[22] Prabhakaran, S. (2020, October 11). Cosine Similarity - Understanding the math and how it works? (with python). ML+. https://www.machinelearningplus.com/nlp/cosine-similarity/

[23] Farha, Ibrahim Abu, and Walid Magdy. "Mazajak: An online Arabic sentiment analyser." Proceedings of the Fourth Arabic Natural Language Processing Workshop. 2019.

[24] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

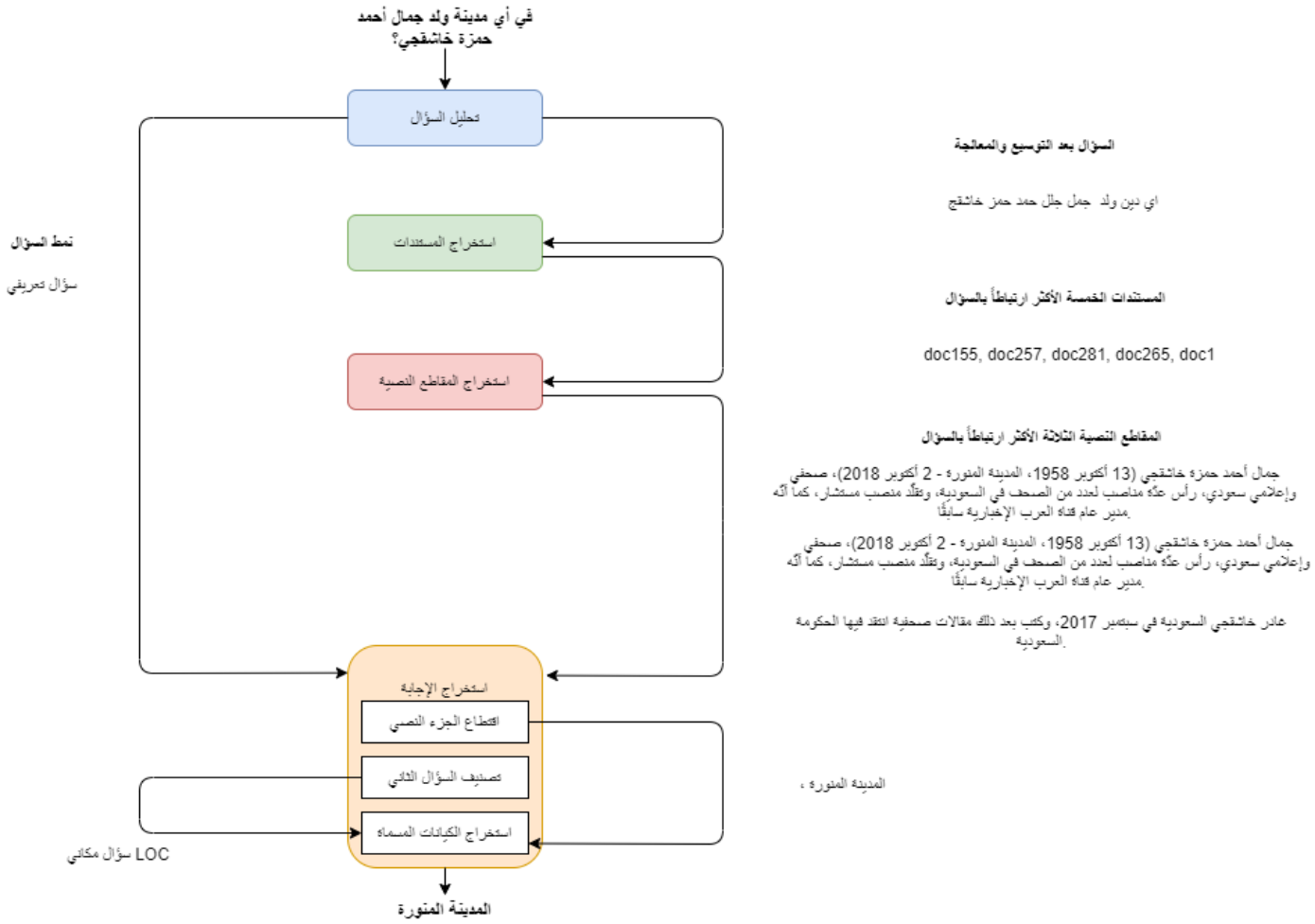
الملحق(1): أمثلة من عينات الاختبار

1.مثال عن إجابة النظام على أسئلة التأكيد



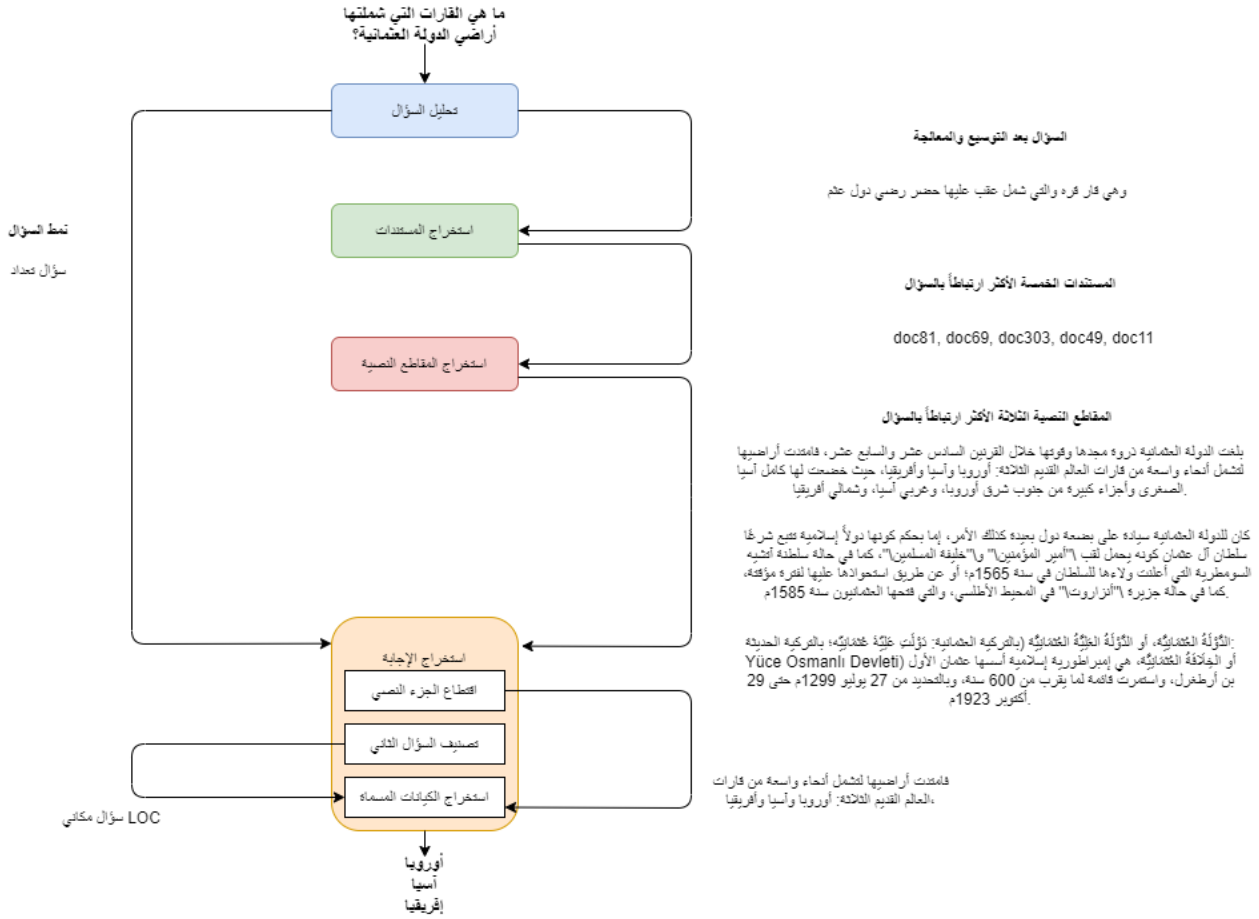
الشكل 35 – مثال عن إجابة النظام المقترح على أسئلة التأكيد.

2.مثال عن إجابة النظام على الأسئلة التعريفية



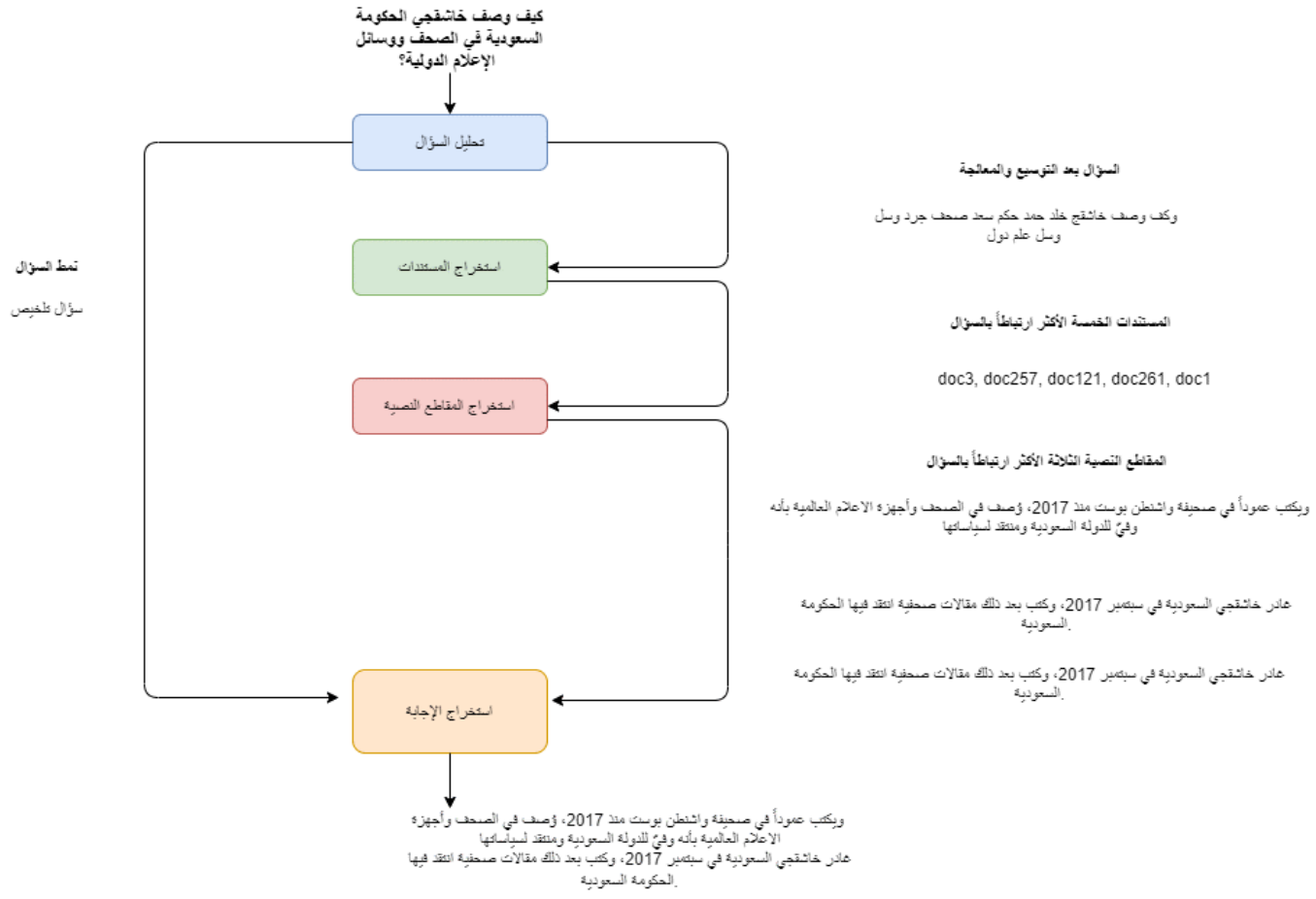
الشكل 36 – مثال عن إجابة النظام المقترح على الأسئلة التعريفية.

3.مثال عن إجابة النظام على أسئلة التعداد



الشكل 37 – مثال عن إجابة النظام المقترح على أسئلة التعداد.

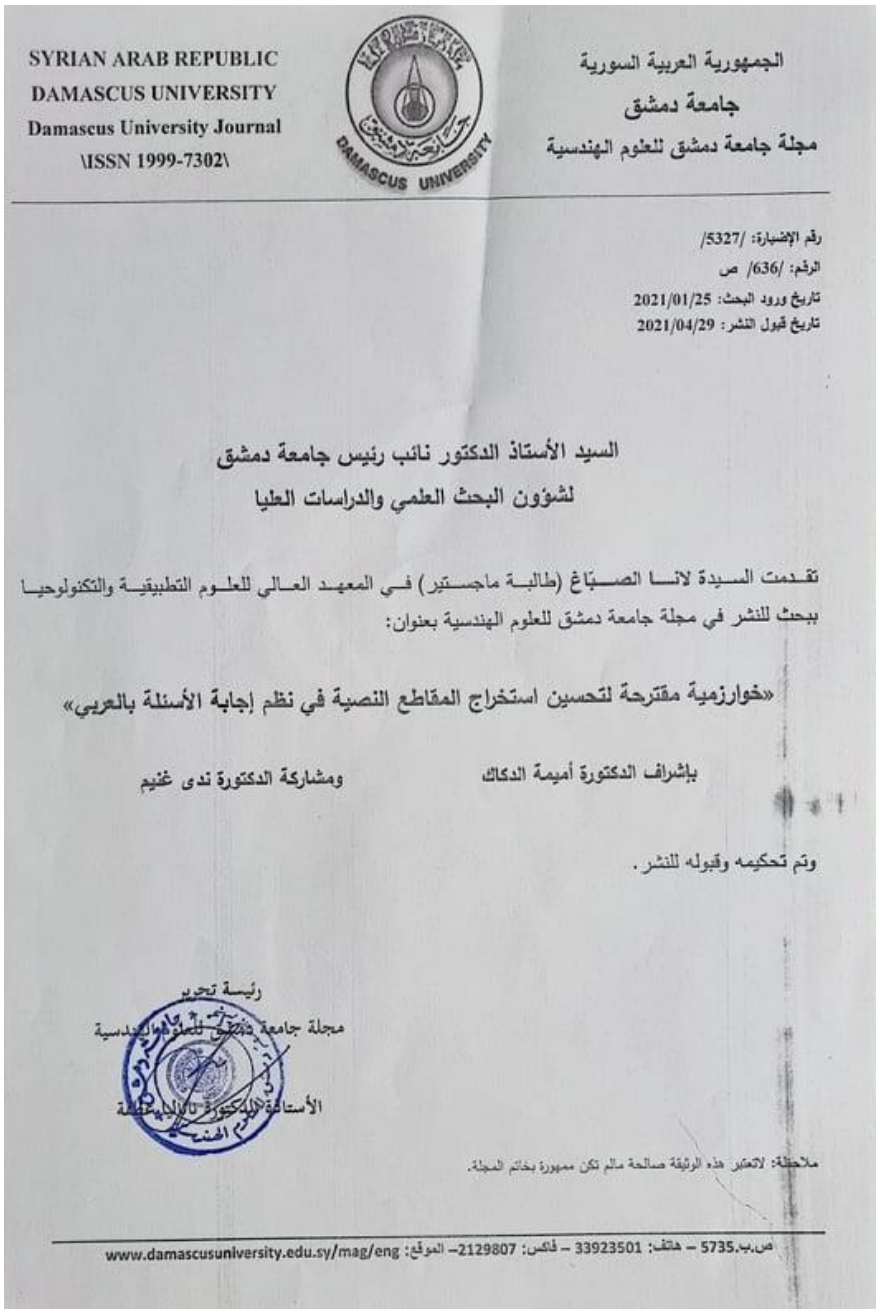
4.مثال عن إجابة النظام على أسئلة التلخيص



الشكل 38 – مثال عن إجابة النظام المقترح على أسئلة التلخيص.

الملحق(2): النشرة المرتبطة بالبحث

جرى تجهيز ورقة بحثية عربية بعنوان "خوارزمية مقترحة لتحسين استخراج المقاطع النصيّة في نظم إجابة الأسئلة بالعربي" وحازت على قبول في مجلة "جامعة دمشق للعلوم الهندسية" كما هو مبين في الشكل(39).



SYRIAN ARAB REPUBLIC
DAMASCUS UNIVERSITY
Damascus University Journal
\ISSN 1999-7302\

الجمهورية العربية السورية
جامعة دمشق
مجلة جامعة دمشق للعلوم الهندسية

رقم الإضبارة: /5327/
الرقم: /636/ ص
تاريخ ورود البحث: 2021/01/25
تاريخ قبول النشر: 2021/04/29

السيد الأستاذ الدكتور نائب رئيس جامعة دمشق
لشؤون البحث العلمي والدراسات العليا

تقدمت السيدة لانا الصبّاغ (طالبة ماجستير) في المعهد العالي للعلوم التطبيقية والتكنولوجيا ببحث للنشر في مجلة جامعة دمشق للعلوم الهندسية بعنوان:

«خوارزمية مقترحة لتحسين استخراج المقاطع النصية في نظم إجابة الأسئلة بالعربي»

بإشراف الدكتورة أميمة الدكاك ومشاركة الدكتورة ندى غنيم

وتم تحكيمه وقبوله للنشر.

رئيسة تحرير
مجلة جامعة دمشق للعلوم الهندسية
الأستاذة الدكتورة [illegible]

ملاحظة: لاتعتبر هذه الوثيقة صالحة مالم تكن ممهورة بخاتم المجلة.

ص.ب.5735 – هاتف: 33923501 – فاكس: 2129807– الموقع: www.damascusuniversity.edu.sy/mag/eng

الشكل 39 – مرفق قبول الورقة البحثية العربية.