تصميم نظام استخلاص المعلومات من بعض فقرات النصوص العربية

بحث مستل من أطروحة دكتوراه

م.د علياء سلمان صابر م.د.سلمى عبد الباقي محمود أ.م.د علي فاضل مرهون

جـامعة البـــصرة

كليـــة العـلوم

قسـم علوم الحاسبات

الملخص

يقدم هذا البحث تصميم نظام حاسوبي وبنائهِ لاستخلاص المعلومات من فقرات مختارة من النصوص الإخبارية المكتوبة باللغة العربية لتوضع في جداول مبوبة لقاعدة البيانات . تتضمن قاعدة البيانات تلك مجموعة من الحقول يختلف عددها ونوعها اعتماداً على مجال عمل النظام والمعلومات المراد استخلاصها من النصوص لغرض تحويل تلك النصوص من تمثيلها النصي غير المهيكل الى تمثيل مهيكل يسهل التعامل معه وأجراء العمليات عليه. لقد تم تصميم النظام المقترح وبناؤه وفق مبادئ وأسس نظرية هندسة البرامجيات ، إذ يتكون من جزأين ، الأول هيكل البيانات Data Structure والمتمثل بالمعجم الساند للنظام والجزء الثاني هو هيكل السيطرة Control Structure والمتمثل بمراحل عمل النظام وهي تحليل المفردات ، والبحث المعجمي ، والتحليل الصرفي ، والتحليل القواعدي_الدلالي ، وتحديد مجال عمل النص واستخلاص الحوادث ، وتنظيم وملء قاعدة بيانات المخرجات ، ومرحلة تقويم مخرجات النظام ، ولإثبات كفاءة النظام المقترح طبق النظام على 75 نصاً من ضمن مجال الحوادث الغير الطبيعية مثل الحوادث الارهابية التي تحدث في بلدنا العزيز مثل الانفجارات وعمليات الخطف ، وكذلك طبق النظام على 50 نصاً من ضمن مجال الحوادث الطبيعية مثل الأعاصير والزلازل ، فضلاً على تطبيق النظام على مجموعة من النصوص خارج نطاق مجالي عمل النظام لأختبار قدرة النظام على تحديد وفرز النصوص .

1- المقدمة

بزيادة كمية المعلومات المتوافرة بشكل نصيّ Textual Information على شبكة المعلومات الدولية أو على هياءة تقارير أو نصوص إخبارية ازدادت أهمية تطوير إمكانيات الحواسيب وقدرتها على التنقيب عن تلك المعلومات بوساطة تطوير آليات التنقيب في الوثائق Document Mining – DM [1] أو التنقيب في النصوص Text Mining – TM من أجل إيجاد المعلومات المهمة أو النماذج المفيدة من بين تلك المعلومات النصية غير المهيكلة Unstructured Textual Information، وعملية التنقيب عن المعلومات تُنجز من خلال العديد من الآليات مثل استخلاص المعلومات Information Extraction – IE [2]استرجاع المعلومات Information Retrieval - IR [3] أو أحيانا تسمى أنظمة استرجاع النصوص Text Retrieval System [4] وتلخيص النصوص Text Summarization – TS [5] وتصنيف النصوص Text Categorization-TC[6] ، وإنّ كلاً من مصطلحات تنقيب الوثائق أو تنقيب النصوص أو اكتشاف المعرفة من قواعد البيانات

النصية Knowledge Discovery in Textual Databases - KDT تُستعمل وبشكل متبادل للتعبير عن مجال العمل نفسه وهو معالجة الوثائق Document Processing [7].

هنالك عدة فوائد من عملية تحويل النصوص من هيأتها غير المهيكلة إلى هيأة أطار مهيكل أي من الشكل النصّي 'Textual Form' إلى الشكل العمليّ أو الجدولي 'Factual or Table form' منها:

- سهولة التعامل أو المعالجة، فعلى سبيل المثال في حالة الوثائق النصية تكون العملية الوحيدة الممكن أجراؤها هو البحث من خلال الكلمات المفتاحية Keyword Word Search ، بالمقابل هنالك مجال واسع من العمليات التي من الممكن أن تجرى على الجداول مثل سهولة العرض ، والبحث عن الحقول ، وترتيب قيود الجدول أبجدياً ، على الاسم أو التاريخ .... ، أو إجراء العمليات الحسابية ، وغيرها من العمليات .

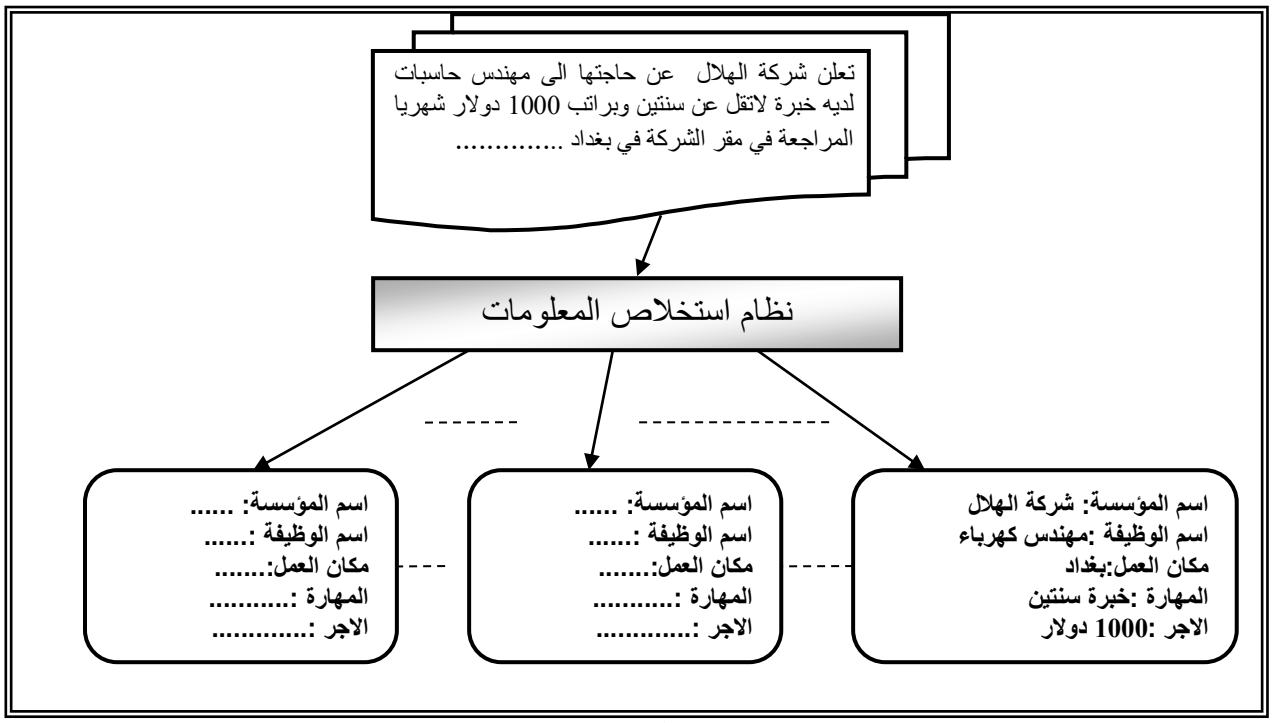
- الفائدة الأخرى من عملية التحويل هو أن الجداول تمثل الفهرس الدلالي للنص Semantic Index ، أي الفهرسة على المعنى بدلاً عن الفهرسة بأعتماد الكلمات المفتاحية كما هو متعارف عليه ، إن الفهرسة المعتمدة على الدلالة أو المعنى هي طريقة لفرض أو تطبيق ترتيب على المعلومات النصية ، إنَّ آلية استخلاص المعلومات هي الطريقة لتطبيق هذا الترتيب آليا .

والشكل (1) يوضح بشكل موجز المهام التي ينجزها نظام استخلاص المعلومات ، في هذا المثال يتركز مضمون الوثائق على إعلانات للوظائف الشاغرة ، المعلومات الهدف أو المراد استخلاصها تُعَّرف على هيأة خمس حقائق مثلاً : اسم المؤسسة التي لديها الوظيفة الشاغرة ،والوظيفة الشاغرة ، والمكان ، المهارة ،والأجر ، .... وغيرها من المعلومات الهدف المراد استخلاصها من النصّ .

وعلى مدى فترة انعقاد ملتقيات فهم الرسائل من الملتقى الأول MUC-1 إلى الملتقى السابع MUC-7 والعديد من أنظمة استخلاص المعلومات قد صمّمت وقوّمت ، وعلى الرّغم من توقف انعقاد تلك الملتقيات بعد الملتقى السابع عام 1991 لأسباب التمويل المادي ما زال عطاء الباحثين في مجال التصميم وتطوير آليات ومهام استخلاص أنظمة استخلاص المعلومات في أوجه متواصلاً معتمدين على الركائز التي أسّستها تلك الملتقيات ، ولكن معظم تلك الأنظمة جاءت لتعالج النصوص المكتوبة باللغة الإنكليزية إلى جانب أنظمة أخرى جاءت لمعالجة النصوص المكتوبة ببعض اللغات الآسيوية الأخرى مثل اللغة الصينية واللغة اليابانية واللغة الكورية ،واهم ما تمتاز به تلك اللغات انه لا يمكن الفصل بين كلماتها من خلال وجود الفراغ كما هي الحال في اللغة الإنكليزية واللغة العربية ، وهذه تعد واحدة من المشكلات التي تواجه بناء أنظمة استخلاص المعلومات لتلك اللغات ، إذا لابد من دعم تلك الأنظمة بمعالجات أولية في مرحلة تحليل المفردات لإنجاز هذه المهمة[8] .

2- نظام استخلاص المعلومات المقترح

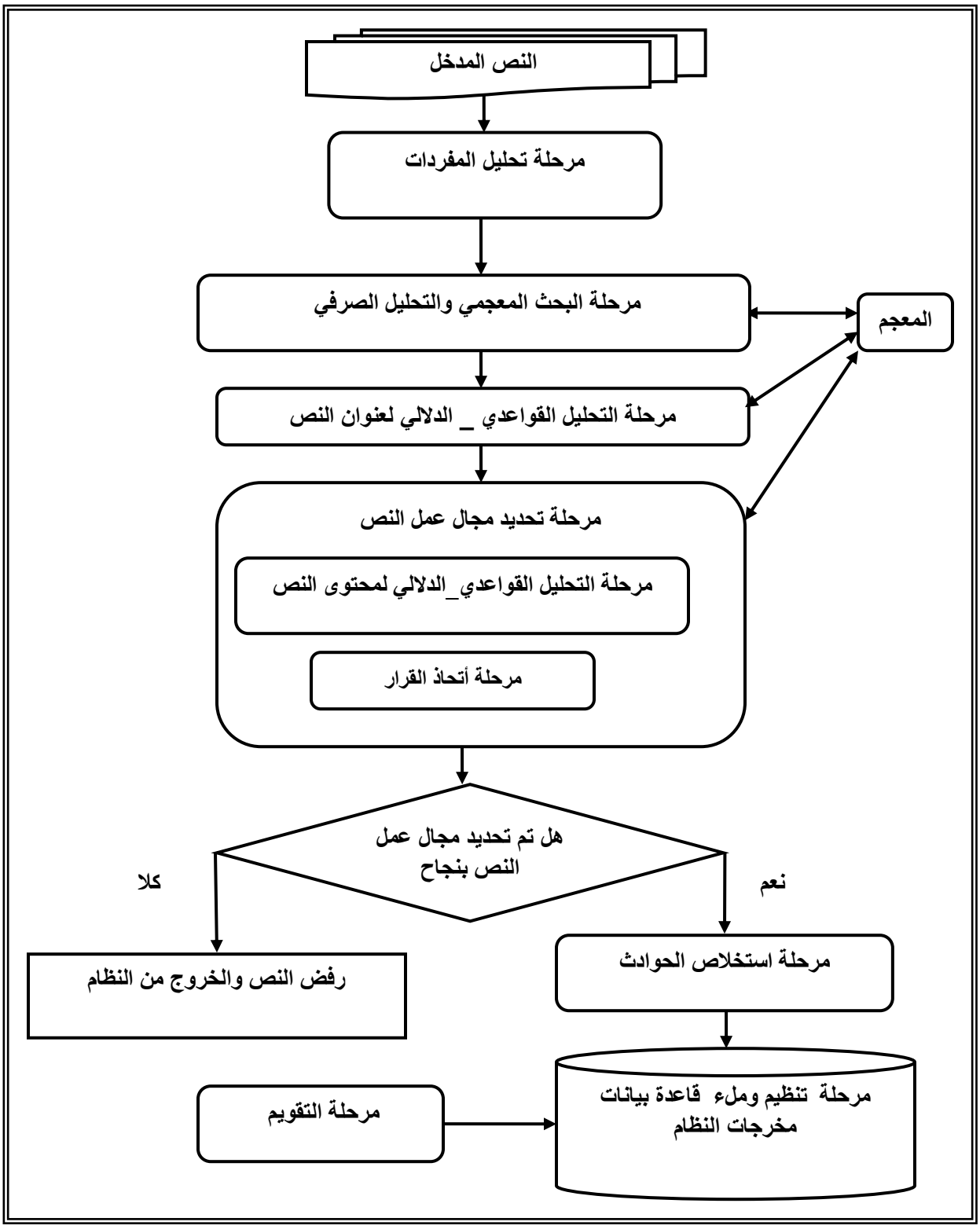
تم تصميم نظام حاسوبي لأستخلاص المعلومات من فقرات مختارة من النصوص الاخبارية لتوضع في جداول مبوبة ، غالبا ما تكون قاعدة بيانات [14] ، والشكل (2) يبين وصفاً لآلية عمل نظام استخلاص المعلومات المقترح وتوضيح اهم مراحل عمله ، ابتداءً من وصف وبيان خصائص النصوص التي يعالجها النظام والية فصلها وتقطيعها الى كلمات في مرحلة تحليل المفردات مرورا بالبحث المعجمي والتحليل الصرفي تمهيدا للتحليل القواعدي _الدلالي وصولاً الى مرحلة تحديد مجال عمل النص لتنتهي بمرحلة استخلاص الحقائق التي تغذي قاعدة بيانات مخرجات النظام . وقد تم اتباع مبادئ وأسس هندسة البرامجيات في تصميم وبناء النظام المقترح المتكون من جزأين هما : هيكل البيانات Data structure والمتمثل بالمعجم الساند للنظام والجزء الثاني هو هيكل السيطرة Control structure والمتمثل بالأليات ومراحل عمل النظام المقترح ، ويندرج هذا النظام ضمن طريقة الهندسة



الشكل(1) مهام نظام استخلاص المعلومات

المعرفية لما تتطلبه من وجود شخص ذي معرفة بمجال عمل النظام من أجل اجراء عملية التحديث والتعديل على انماط الاستخلاص ، وقبل الولوج في تفاصيل العمل كل مرحلة من مراحل عمل النظام لابد من ان نبين انه قد طُبق النظام على 75 نصاً من ضمن مجال العمل الأول المتمثل بالحوادث الارهابية الغير الطبيعية التي تحدث في بلدنا العزيز العراق والناجمة من عمليات التفجير والقتل والخطف ، أما في مجال العمل الثاني فقد طبُق على 50 نصاً من ضمن مجال الحوادث والكوارث الطبيعية التي قد تحدث في العالم مثل الأعاصير والفيضانات والزلازل ، فضلا على مجموعة من النصوص خارج نطاق مجالي عمل النظام لبيان أو اختبار قدرة النظام على تحديد وفرز النصوص التي ليست ضمن مجالي عمله ، وفيما يلي شرح موجز لكل جزء من اجزاء النظام واهميته في انجاز عملية الاستخلاص [14].

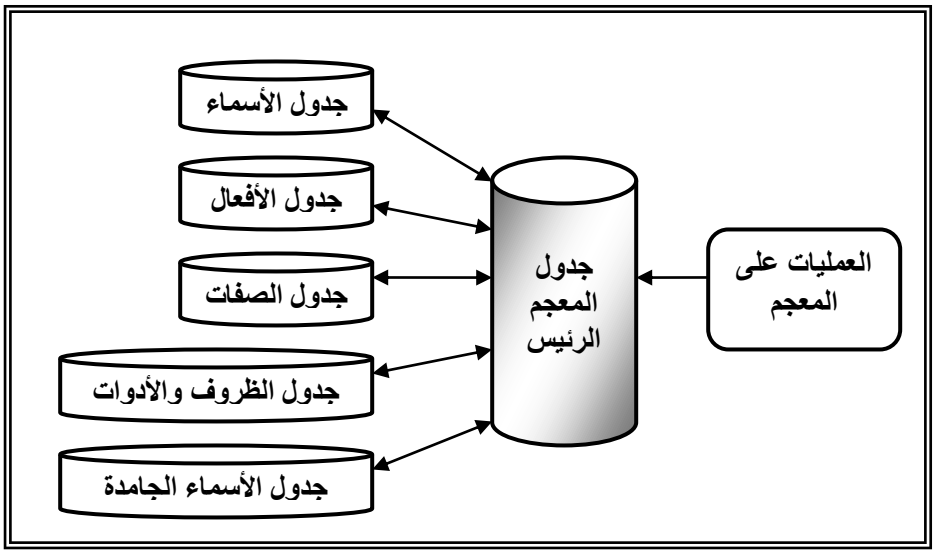
- **النص المدخل :** لقد أعُتمد في تصميم واجهة إدخال النصوص على توفير المرونة للمستخدم في إدخال النصّ المراد العمل عليه ، وتعديلهُ ، وخزنهُ ، وطباعته ، إذ يمكن إدخال النص المراد معالجته مباشرةً أو أن يُستردّ من الموقع الخزني الذي خُزن فيه، وأن جميع النصوص التي عُولجت سواءً كانت من ضمن مجال الحوادث الطبيعية أو مجال الحوادث غير الطبيعية اتّخذت بوصفها عّينات من النصوص الإخبارية المنشورة في الصحف أو المواقع الالكترونية على شبكة الانترنت مثل موقع www.aljazeera.net و www.bbcarabic.com وغيرها من المواقع الإخبارية .



الشكل (2) المخطط العام للنظام المقترح

- **المعجم :** المعجم هو مستودع للمفردات اللغوية وسماتها وخصائصها المعجمية والنحوية والدلالية [9]، صُمّم المعجم ليتضمن أصناف الكلمات في اللغة العربية وهي (الاسم ،والفعل ،والصفة ، والأدوات

والظروف ،فضلاً عن الأسماء الجامدة)، ونظراً لاختلاف الخصائص التركيبية والخصائص المعجمية لكل صنف من هذه الأصناف قد اتُبعت مفاهيم نظرية الحقـول الدلالية لتصنيف مفردات المعجم المقترح ، التي تعتمد على فكرة أعطاء مفردات اللغة شكلا . تتكون منظومة المعجم المقترح والموضحة بالشكل (3) من الجدول الرئيس الذي يتكون من حقلين الأول يمثل الجذع للمفردة والثاني تخزن فيه رقم يشير إلى صنف المدخل وكالآتي ("1" اسم ،"2" فعل ، "3" الأدوات والظروف ، "4" الصفات ، "5" الأسماء الجامدة) ، والجدير بالإشارة ، أن بعض المفردات تكون مشتركة في أكثر من مدخل (أي حالة لبس قواعدي) ، ليتحّقق تأشير وجود حالة لبس قواعدي تكـفُّلاً بحلها لاحقا إمّا خلال مرحلة التحليل الصرفي أو مرحلة التحليل القواعدي_الدلالي .منظومة المعجم المقترح تتكون من ستة جداول ، الجدول الرئيس وخمسة جداول لأصناف الكلمات التي ترتبط معه كل جدول يتضمن عدداً من الحقول الذي قد يختلف عددها من جدول لأخر لحفظ المعلومات المعجمية والقواعدية والدلالية الخاصة بكل صنف. يمكن أن نستدل أن العمليات التي يمكن أجراؤها على مفردات المعجم تجري تحت إدارة و إشراف الجدول الرئيس .



الشكل(3) منظومة المعجم المقترح

- **مرحلة تحليل المفردات :أن** المهمة المعتادة لمحلل المفردات في معظم تطبيقات معالجة اللغات الطبيعية هو تقطيع الجمل المدخلة إلى كلمات منفصلة ، وفي نظامنا المقترح أوكلنا مهاماً أضافية لعمل محلل المفردات تتلخص بمهمة فصل عنوان النص عن محتواه وكذلك مهمة التدقيق الإملائي .
- **مرحلة البحث المعجمي والتحليل الصرفي:** بعد أن توضّح هيكلية منظومة المعجم وخصائصه ، يأتي الدور لتوضيح مهام مرحلة البحث المعجمي ، إذ إنَّ مدخلات هذه المرحلة هي قائمتي مفردات عنوان ومحتوى النصّ . تنفذ عملية البحث عن كل كلمات تلك القائمتين ، في جدول المعجم الرئيس . في حالة وجودها يُتوجْه إلى جدول الصنف المحدد أو الأصناف في حالة الاشتراك لجلب الخصائص والمعلومات المخزونة مع الكلمة ،أما مرحلة التحليل الصرفي فتبدأ بعد فشل مرحلة البحث المعجمي ، إذ تجرى عملية التحليل الصرفي للكلمة (اسم، فعل ، صفة) وذلك بفصلها إلى جذع ولواصق (سوابق ولواحق) مع تحديد الخصائص المعجمية لجذع الكلمة المتصرفة ولجميع إضافاتها ، ويستند المحلل الصرفي نظاماً تحليلياً إلى مكونين أساسيين وهما :

1- **المكونات الوصفية** : تتمثل أولا بالمعجم الذي وضّحت خصائصه وطريقة تنظيمه في الفقرات السابقة ، وثانيا استُثمرت فكرة شبكة الانتقالات المعززة آليةً ناجحةً لتحليل الكلمة إلى سوابق ولواحق وجذع . يكمن سر النجاح في استعمال قواعد شبكة الانتقالات المعززة في أنظمة معالجة اللغات الطبيعية لكونها أداة تفسير وتحليل ناجحة ، إذ صُممت شبكة انتقالات معززة محددة [10] لتمثيل المعرفة الحساسة السياق حول العلاقة ما بين الجذع واللواصق الصرفية.

2- **المكونات الإجرائية :** تتمثّل باقتراح آلية بحث العمق أولاً طريقة لاجتياز الشبكة من اجل الوصول إلى الحل الأمثل المتمثل بتحليل صرفي صحيح واحد للكلمة المراد تحليلها.

- **مرحلة التحليل القواعدي _الدلالي :** تعد مرحلة التحليل القواعدي أو عملية الإعراب أحدى الأركان الأساسية من ضمن مراحل عمل النظام إذ تعتمد مراحل النظام اللاحقة وصحتها على صحة ودقة نتائج هذه المرحلة ، والهدف من هذه المرحلة هو تقطيع الجمل الفرعية إلى تراكيب أو جمل نحوية اصغر قد تكون اسميةً ،أو فعليةً،أو جاراً ومجروراً ،أو ظرفية .سنصطلح على تسميتها الأنماط القواعدية Syntactic Pattern . إن الجمل الفرعية قد تحتوي على بعض أدوات وحروف النسخ والنصب والجزم وأسماء الإشارة والأسماء الموصولة التي يُبقى عليها بوصفها نماذج منفصلة تسبق أو تربط الأنماط القواعدية . اعتُمدت آلية الإعراب من أسفل_إلى _أعلى كإطار عام لعملية التحليل التي تعتمد على فكرة البحث من ضمن الجملة المراد تحليلها على الأنماط القواعدية البسيطة أي أجراء عملية مطابقة مع أنماط نحوية معرفة مسبقا ليتحقق دمجها أو توحيدها لاحقا لتكوين أنماط قواعدية مركبة وفي نظامنا المقترح يمكن أن تجرى عملية الدمج لتلبية المتطلبات **النحوية** أو **السياقية** وهي من مهام هذه المرحلة أو لمتطلبات **دلالية** وهي من مهام مرحلة تحديد الأصناف الدلالية والدمج الدلالي ،و بعد تجاوز مرحلة التحليل القواعدي بنجاح يتحقِّق الانتقال إلى مرحلة التحليل الدلالي التي تهدف الى تحديد الأصناف الدلالية للأنماط القواعدية البسيطة والمركبة التي حُدّدت وكُوّنت في المرحلة السابقة ، فالصنف الدلالي يعطى لكل نمط تبعا للمعنى العام الذي يمكن أن تأخذه تلك الجملة التي تحدّد من خلال تطبيق مجموعة من الشروط الانتقائية وعمليات الفحص والمطابقة على خاصية الإطار أو السمة الدلالية مع خاصية شفرات الكيانات التي قد ترتبط مع الأسماء أو الأفعال ، فبعض الأصناف تبقى ملازمة للنمط ولا تتغير بعد تحديدها وبعضها الآخر قد يتغير نتيجة لعملية الدمج الدلالي أو نتيجة المعالجات السياقية للكلمات المؤثرة في النمط ، ويمكن أن نوضح المزيد من أسس تحديد الأصناف الدلالية بالأتي :

- **صنف النمط القواعدي** ، إنَّ للصنف نمط الجملة (اسمية ، فعلية ، جار ومجرور أو ظرفية) التأثير المباشر على صنفها الدلالي.
- **مجال عمل النظام** ، حيث لمجال عمل النظام التأثير غير المباشر على ما يمكن أن تأخذه الأنماط من أصناف دلالية وذلك لتأثيره في السياقات والمعاني التي تأخذها الكلمات المؤثرة في النمط القواعدي وبالتالي في الصنف الدلالي للنمط.

- **قاعدة بيانات المخرجات** :أعطيت مسميات للأصناف الدلالية بطريقة تساعد على تحديد ملامح الأحداث أو الحقائق المراد استخلاصها من ضمن مجالي عمل النظام وذلك من خلال ارتباط دلالاتها بشكل مباشر أو غير مباشر مع حقول قاعدة بيانات المخرجات التي يهدف النظام إلى تكوينها من استخلاص المعلومات من النصوص ، و الجدول (1) يوضح حقول قاعدة بيانات مخرجات النظام المقترح الشاملة والمتكونة من تسعة عشر حقلا التي قد تشترك مع بعضها آو تختلف تبعا للخصائص الآتية :
  - **قياسي** : تشير هذه الخاصية إلى الحقول التي اعتمدت من قاعدة بيانات مخرجات الأنظمة المشاركة ملتقى النصوص الثالث والرابع [11 ،12 ، 13]

- **غير قياسي** : هي خاصية الحقول التي أضُيفت لتتلاءم مع متطلبات مخرجات مجالي عمل النظام المقترح .
- **مشترك** : أما خاصية الحقل المشترك فتعطى للحقول المشتركة بين المجالين .
- **خاص بالمجال** :تخصص هذه الخاصية لبعض الحقول التي تكون خاصة بمجال عمل من دون الآخر ، إذ يكون لهذه الخاصية دور مهم في مرحلة تحديد مجال عمل النظام .
- **متطلبات النظام** :وتضاف هذه الخاصية لبعض الحقول التي أُضيفتْ تلبية لمتطلبات عمل مراحل النظام .

الجدول (1) حقول قاعدة البيانات مخرجات النظام الشاملة

ت	اسم الحقل	قياسي	غير قياسي	مشترك	خاص بالمجال	متطلبات النظام
1	اسم النص			✓		✓
2	عدد الجمل			✓		✓
3	رقم الجملة			✓		✓
4	مجال العمل		✓	✓		✓
5	سبب الحادث الرئيس	✓		✓		
6	سبب الحادث الثانوي		✓	✓		
7	سبب أو نتيجة الحادث		✓		2	
8	سلاح الحادث	✓			1	
9	المتسبب في الحادث	✓			1	
10	الزمان	✓		✓		
11	المكان	✓		✓		
12	التأثيرات المميتة			✓		
13	التأثيرات غير المميتة		✓	✓		
14	التأثير على الأهداف الفيزيائية	✓		✓		
15	الخطف		✓		1	
16	الفقدان		✓		2	
17	ترك مكان الإقامة		✓		2	
18	مصدر المعلومات	✓		✓		
19	مقياس تأثير الحادث		✓		2	

ملاحظة :(1 : المجال الاول ، 2: المجال الثاني ، ✓ مشترك بين المجالين)

- **مرحلة تحديد مجال عمل النص** : الهدف من هذه المرحلة هو التحقق من عائدية النص المدخل أو تصنيفه إلى احد مجالي عمل النظام دون الآخر عندها يُقْبلُ النص لينتقل عمل النظام إلى مرحلة الاستخلاص أو أن يُرْفض النص في حالة عدم أثبات عائديته إلى أحد المجالين . ومن الشكل (1) الذي يبين المخطط العام لعمل النظام ، يمكننا أن نستدل على أن مرحلة تحديد مجال عمل النظام تعتمد بشكل كبير على نتيجة التحليل القواعدي_الدلالي لعنوان النص ولكن هذا لا يعني أن عملية اتخاذ قرار قبول النص أو رفضه تعتمد فقط على نتائج هذا التحليل بل يُعْتمّد أيضا على نتيجة التحليل القواعدي_الدلالي لمحتوى النص للتعزيز من صحة القرار المتخذ إذ تجري عملية التحليل القواعدي الدلالي لمحتوى النص ضمنيا مع مراحل عمل هذه المرحلة .تنجز هذه المرحلة مهامها من خلال أجراء سلسلة من عمليات البحث والتدقيق على شفرات أصناف الاستخلاص والأصناف الدلالية للأنماط القواعدية لجمل عنوان النص ومحتواه من أجل التحقق من وجود أركان الحدث المراد استخلاصه المتمثلة بسبب الحادث الرئيس أو نتيجة الحادث أو ما يمكن أن نسميه أيضا السبب الثانوي للحادث فضلاً عن زمان ومكان الحادث وكذلك التحقق من وجود شفرات لأصناف استخلاص ترتبط بالحقول الخاصة بمجال من دون الأخر من اجل دعم آلية اتخاذ القرار . وقبل الدخول في تفاصيل آلية عمل هذه المرحلة لابد من توضيح ماذا نعني بأصناف أسباب الحوادث الرئيسة والثانوية التي ترتبط مع مجالي عمل النظام الذي يبين أن الأسباب الرئيسة في مجال الكوارث الطبيعية هي (الزلازل ،الأعاصير ، الهزات الأرضية ، السيول ، الفيضانات) أما الحوادث التي يمكن أن

تكون سبباً أو نتيجةً للحوادث الطبيعية فهي (الأمطار الغزيرة ، ذوبان الثلوج ، الانهيارات الأرضية ، الانجرافات الأرضية) .

وفي آلية اتخاذ القرار فيتم إنهاء سلسلة الاختبارات التي تسلكها المسارات الفرعية للوصول إلى قرار أما بقبول النص وتحديد مجال عمل النظام لتحدد عندها قيمة الحقل Domain_Type من ضمن قيد التحليل وكذلك تعطى درجة تمثل مستوى نجاح عملية اتخاذ القرار بوصفه قبولاً جيداً أو قبول متوسط أو قبول ضعيف إذ تحدّد بالاعتماد على نوع المسار الفرعي الذي يُجتاز للوصول إلى القرار ، أو إنَّ يُرفض النص المدخل في حالة فشله في اجتياز أي من مسارات الاختبارات الفرعية .

- **مرحلة استخلاص الحوادث :** الهدف من هذه المرحلة هو استخلاص المعلومات من النصوص التي أخُذ القرار بقبولها وبغض النظر عن درجة القبول ليخزن المعلومات المستخلصة في قاعدة البيانات الشاملة كمرحلة أولى والموضحة مسبقا في الجدول (1) ليتم فرزها بعد ذلك إلى جدولين ، الأول خاص بنتائج مجال الحوادث غير الطبيعيَّة والمتكون من أحد عشر حقلا والثاني خاص بنتائج مجال الكوارث الطبيعية والمتكون من اثني عشر حقلا والموضح تفاصيلهما في الجدول (2) ، تبدأ آلية الاستخلاص المقترحة مهامها بعملية فحص للمعلومات التي تتضمنها قيود التحليل لجمل محتوى النص ، إذ تتركّز بشكل رئيس أما على خاصية صنف الاستخلاص فقط أو على خاصيتي صنف الاستخلاص والصنف الدلالي معا وبشكل ثانوي يُركّز على صنف النمط القواعدي.لقد صُمِّم نهجا معالجة لإنجاز مهام هذه المرحلة ،الأول لاستخلاص المعلومات من الأنماط القواعدية التي لها أصناف استخلاص ترتبط بشكل مباشر مع بعض حقول قاعدة البيانات الشاملة،والثاني لاستخلاص المعلومات من الأنماط القواعدية التي لها ارتباط غير مباشر مع بعض حقول قاعدة البيانات الشاملة.

الجدول (2) حقول جدولي مخرجات مجالي عمل النظام

جدول حقول المجال الأول

ت	اسم الحقل
1	سبب الحادث الرئيس
2	سبب الحادث الثانوي
3	سلاح الحادث
4	التأثيرات المميتة
5	التأثيرات غير المميتة
6	الزمان
7	المتسبب في الحادث
8	التأثيرات الفيزيائية
9	الخطف
10	المكان
11	مصدر المعلومات

جدول حقول المجال الثاني

ت	اسم الحقل
1	سبب الحادث الرئيس
2	سبب الحادث الثانوي
3	سبب أو نتيجة الحادث
4	التأثيرات المميته
5	التأثيرات الغير مميتة
6	التأثيرات الفيزيائية
7	الفقدان
8	مصدر المعلومات
9	الزمان
10	المكان
11	ترك مكان الإقامة
12	مقياس الحادث

- **تنظيم مخرجات النظام:** بعد أن تنهي آلية الاستخلاص مهامها بخزن المعلومات التي استخُلصت من الأنماط القواعدية ذات أصناف الاستخلاص سواءً كانت مرتبطة بشكل مباشر أم غير مباشر بحقول قاعدة البيانات الشاملة في قائمة الخزن المؤقتة المعدة لهذا الغرض لتجرى عليها عمليات تعديل مثل حذف المعلومات المكررة من بعض الحقول مثل سبب الحادث الرئيس أو سبب الحادث الثانوي وذلك نتيجة لوجود أكثر من نمط له صنف الاستخلاص نفسه ليحصل بعد ذلك نقل المعلومات المستخلصة تخُزّن في قاعدة البيانات الشاملة وثمَّ تُفصل بالاعتماد على حقل صنف المجال إلى جدولين كل جدول يتضمن

مجموعة الحقول المشتركة والخاصة بمجال العمل ومما سهل العمل كثيراً ما توفره أنظمة قواعد البيانات من عمليات تسهل من التعامل والوصول الى البيانات مثل البحث والترتيب وغيرها من العمليات ،ومن خلال الجدول (3) والجدول (4) اللذين يوضحان نتائج تطبيق نتائج آلية الاستخلاص على محتوى النص في الملف Text1 والنص في الملفText107 اللذين تم الاستعانة بهما لتوضيح مراحل عمل النظام جميعها .

استشهاد اثنا عشر شخصاً في انفجار سيارة ملغومة في بغداد اليوم الخميس

انفجرت سيارة ملغومة بالقرب من مركز للشرطة في العاصمة بغداد ، وقالت التقارير ان اثني عشر شخصاً استشهدوا في الأنفجار وأصُيب عدة اخرون .

الجدول (3) نتائج استخلاص المعلومات من النص في الملف text1 (المجال الاول)

زلزال بقوة 6.5 درجة على مقياس ريختر يضرب جزيرة سومطرة

ضرب زلزال بقوة 6.5 درجة على مقياس ريختر جزيرة سومطرة وتسبب في قتل العشرات واصابة المئات

اسم الحقل	المعلومات المستخلصة
سبب الحادث الرئيس	هجوم بالمتفجرات
سبب الحادث الثانوي	تأثيرات مميته + تأثيرات غير مميتة
سلاح الحادث	سيارة ملغومة
التأثيرات المميتة	اثنا عشر شخصاً
التأثيرات غير المميتة	عدة اخرون
الزمان	-
المتسبب بالحادث	-
التأثيرات الفيزيائية	-
الخطف	-
المكان	القرب من مركز شرطة في العاصمة بغداد
مصدر المعلومات	التقارير

وتدمير العديد من المباني.

أسم الحقل	المعلومات المستخلصة
سبب الحادث الرئيس	الكوارث الطبيعية_زلزال
سبب الحادث الثانوي	تأثيرات مميته+تأثيرات غير مميتة +تأثير على الأهداف الفيزيائية
سبب أو نتيجة الحادث	-
التأثيرات المميتة	العشرات
التأثيرات الغير مميتة	المئات
التأثيرات الفيزيائية	تدمير العديد من المباني
الفدان	-
مصدر المعلومات	-
الزمان	-
المكان	جزيرة سومطرة
ترك مكان الأقلمة	-
مقياس الحادث	6.5 درجة على مقياس ريختر

الجدول (4) نتائج استخلاص المعلومات من النص في الملف text107 (المجال الثاني)

مرحلة التقييم : تخضع النتائج التي يُحْصَلُ عليها من تطبيق نظم استخلاص المعلومات إلى آليات التقويم وطرائقه على وفق المقاييس التي ذُكِرت في الفصل الأول وهي مقياس الاسترداد Recall لإيجاد نسبة المعلومات المستخلصة بشكل صحيح من النص إلى المعلومات ذات العلاقة الموجودة من ضمن النص ومقياس الدقة Precision لإيجاد نسبة المعلومات المستخلصة من النص بوساطة نظام الاستخلاص إلى المعلومات الواجب استخلاصها بشكل صحيح من النص أو مقياس أف F-measure الذي يجمع بين المقياسين السابقين بعلاقة

رياضية تعكس مستوى تلازم المقياسين وأهمّيته ، إن تلك المقاييس عادة ما تنجز من خلال إجراء سلسلة من الاختبارات والمقارنات آليا أو يدويا بين المعلومات التي يُحْصَلُ عليها من التطبيق الأتي للنظام مع ما يسمى نتائج الفحص التي يُحْصَلُ عليها من استخلاص المعلومات من نصوص الفحص يدوياً بوساطة الخبراء أو العاملين على تقويم النظام ، وفيما يخص نظامنا المقترح فقد اعتمد مقياس الدقة آلية لتقويم وفيما يأتي تفسير لأسباب اعتماد هذا المقياس من دون غيره وآلية تطبيقه .

فيعزى السبب الرئيس لاختيار مقياس الدقة من دون غيره لطبيعة النصوص المطبقة على النظام ، فمعظم النصوص متكونة من مجموعة من الفقرات المختارة التي تكون من ضمن مجالي عمل النظام وبمعنى أخر أنَّ آلية عمل النظام لا تعُنى باقتطاع المقاطع أو الفقرات ذات العلاقة بمجال عمل النظام من بين الفقرات الكلية المكونة للوثيقة أو النص لما يتطلبه هذا الأجراء من فهم كامل للنص وهو خارج نطاق أهداف النظام المقترح وهذا جعلنا نستبعد مقياس الاسترداد لأنه يعتمد على إيجاد النسبة بين المعلومات المستخلصة بشكل صحيح إلى المعلومات المحتملة في النص (ذات العلاقة وغيرها) وهذا ما لا يتوافر في طبيعة النصوص المطبقة .

بالرغم من جميع تلك المحددات والمعوقات والصعوبات التي واجهت إنجاز النظام المقترح يمكن أن نصف مستوى إنجاز النظام بالجيدّ وجاء وصفنا هذا معتمد على نتائج مرحلة التقويم ، ومن خلال مشاهدة نتائج المجالين ومن خلال حساب النسبة المئوية للدقة الكلية لنتائج الاستخلاص لكلا المجالين التي حُسِبتْ على وفق المعادلة الآتية :

$$\text{النسبة المئوية للدقة} = \frac{\text{مجموع المعلومات المستخلصة آليا لجميع النصوص}}{\text{المعلومات المستخلصة يدويا لجميع النصوص}} \times 100$$

وجدنا أنَّ النسبة المئوية لدقة نتائج المجال الأول = 89 % ، أما النسبة المئوية لدقة نتائج المجال الثاني=95 % ، فهذا يبين أن النظام يعمل بفعالية أحسن على نصوص مجال الكوارث الطبيعية منه على نصوص مجال الحوادث غير الطبيعية ويمكن أن يُعزى ذلك إلى عدة أسباب منها عدد النصوص وحجمها التي تخص المجال الثاني تكون اقل عددا واصغر حجماً منها للمجال الثاني فضلاً عن طبيعة النصوص من حيث التشابه إلى حد ما في الأنماط القواعدية المكونة لمحتوى النص للنصوص مجال الكوارث الطبيعية ، ومن جانب آخر يمكن أن نصف نسبة دقة المجال الأول بالجيدة وبالمرضية نسبة إلى عدد وحجم النصوص فضلاً عن التغاير الواضح لأنماط الجمل المكونة له .

3- الاستنتاجات :

- تعد طبيعية النصوص المعتمدة أولى المحددات التي واجهت تصميم النظام المقترح من حيث القيود الموضوعة على حجم كتابة عنوان النّص ونمطه وكذلك عدم قدرة النظام على عزل النصوص المكتوبة من الصور التي غالبا ما تكون ملازمة للنصوص الخبرية بشكل آلي.

- وتعد محدودية مفردات المعجم وخصوصيتها من ضمن مجالي عمل النظام وعدم تمكنه من استنباط العلاقات المعجمية المعرفية بين المفردات المدخلة وكذلك عدم توافر معجم معرفي الكتروني موحد يزيل عن كاهل الباحثين عبء بناء معاجم تكون خاصة لأغراضهم البحثية وهو ما قد أضاف قيدا جديدا على فضاء عمل النظام .

- تُعدَّ مرحلة التحليل القواعدي واحدة من أصعب مراحل عمل النظام وأدقّها فتداخل ومرونة تكوين الجمل المكونة للغة العربية وعدم وفرة الأنموذج التمثيلي العام الذي يمكن أن يضم جميع القواعد النحوية التي تتكون منها الجمل ، فضلاً عن ندرة الأعمال والجهود البحثية التي تُعنى بمعالجة النصوص من حيث المعالجة القواعدية واقتصارها على معالجة أنماط محددة من الجمل المنفردة ، من هنا جاءت مرحلة التحليل القواعدي لنظامنا المقترح لتعطي نوعاً من المرونة في تكوين الأنماط القواعدية البسيطة والمركبة المتكونة منها جمل النصوص المطبقة على النظام ومن ضمن أنماط محددة من الجمل الاسمية والفعلية والجار والمجرور والظرفية من خلال تصميم آليتين للحالات المحدودة لاكتشاف الأنماط القواعدية البسيطة للجمل الاسمية والفعلية وكذلك اقتراح إستراتجية إعراب معتمداً على مبدأ الإعراب من أعلى إلى أسفل من أجل تحديد أولويات الأنماط التي يُبْدأ بتكوينها أولا ،أما الأنماط المركبة تُتكوّن من عمليات الدمج النحوي أو الدمج الدلالي ، وبالرغم من النجاح الذي حققته هذه المرحلة تعّد من محددات عمل النظام المقترح لعدم معالجتها الأنماط المعقدة التي قد تتضمن أدوات النفي والجزم والاستفهام والإشارة وغيرها فضلاٍ عن الإخفاق الذي قد يحدث في تحديد النمط القواعدي الصحيح نتيجة لوجود حالات اللبس القواعدي .

- مرحلة الاستخلاص اعتمدت على استخلاص المعلومات من الأنماط القواعدية التي تكون لها ارتباط مباشر أو ارتباط غير مباشر بحقول قاعدة بيانات المخرجات التي قد أعطت نتائج استخلاص يمكن أن نصفها بالجيدة من خلال تفحصنا لنتائج هذه المرحلة. وتوصلنا أيضاً إلى أن هنالك بعض الأصناف تكون احتمالية حدوث أخطاء في أثناء عملية استخلاصها قليلة مثل أصناف الزمان والمكان وأصناف تكون احتمالية الخطأ في استخلاصها متوسطة الحدوث مثل أسباب الحوادث الرئيسة والثانوية أما الأصناف ذات الاحتمالية الأكثر عرضة للوقوع في الأخطاء في أثناء عملية الاستخلاص فهي التي قد لا ترد أصنافاً منفصلة في الجملة بشكل دائم مثل سلاح الحادث أو التي يُحدِّد صنفها الدلالي بشكل خاطئ .

4- المصادر :

1- Dixon .M. ,"An overview of Document mining Technology" , 1997.

2- Cowie .J. and Wills .Y. , "Information Extraction ", The University of Sheffield ,

department of computer Science ,UK, W.D ,2004.

3- Lazarinis ,F. , "Combining Information Retrieval with Information Extraction for Efficient Retrieval of calls for papers" , Dept. Of Computer Science , university Of Glasgow , Glasgow , Scotland ,W.D ,2003.

4- Grishman .R. , " Information extraction :Techniques and challenges " , Computer Science Dept. New York University ,new York , USA. , W.D ,2004.

5- Han .J. , "Mining Knowledge at Multiple Concept Levels " , School of Computing Science , Simon Fraser University , Canada ,W.D,2002.

6- Bagga .A. , " A Short Course on information Extraction :A proposel" ,Dept. of computer science , Duke University ,Durham , 1998.

7- Ba gga .A., Chai .J. And Biermann .A. , "Extraction Information From Text", Dept. Of Computer Science , Duke University , Durham ,W.D,2005.

8- Gariglinao .R. , Urbanowicz .A . And Netteton .D. J. ," University of Durham: Descrption of the LOLITA system As used in MUV-7 ", Laboratory for language engineering , dept. of computer Science , University of Durham , UK , W.D,2003.

9 - علي، د.نبيل ،"اللغة العربية والحاسوب(دراسة بحثية) " ، تعريب للطباعة والنشر، القاهرة ،1988.

10 – د.علياء سلمان صابر ، د.سلمى عبد الباقي محمود ، د.علي فاضل مرهون ، معالج صرفي تحليلي للواصق الكلمة المتصرفة في اللغة العربية باستخدام بحث العمق أولا في البحث في شبكة الانتقالات المعّززة ، مجلة ابحاث البصرة (العلميات) العدد الرابع والثلاثون ، الجزء السادس ،15 كانون الاول 2008 ، الصفحة 1- 9 .

11- Wilks .Y. and Catizone .R., " Can we make Information Extraction More Adaptive? " , The University of Sheffield , department of computer Science ,UK , W.D,2005.

12- Grishman .R., " The NYU System For MUC – 6 Or Where's the Syntax " ,Computer science Dept. , New York University , USA., W.D ,2005.

13- Fukumoto .J. , Masui .F. , Shimohata .M. and Sasaki .M. ," OKI Electric Industry: descripation of the OKI system as Usedd for MUC-7" , Kansai Lab and R&D group , Oki electric industry Co., Japan , W.D,2005.

14- د.علياء سلمان صابر ، " تصميم نظام استخلاص المعلومات من بعض فقرات النصوص العربية "، جامعة البصرة ، كلية العلـــــــوم قســـــــم علـــــــوم الحاسبـــــــات ، اطروحـــــــة دكتـــــــوراه ،2007 .