



الجمهورية العربية السورية

جامعة دمشق

كلية الهندسة الميكانيكية والكهربائية

قسم هندسة الحواسيب والامتة

التلخيص الآلي للنصوص العربية باستخدام الأنطولوجيا

رسالة مقدمة لنيل درجة الدكتوراه في هندسة الحواسيب وشبكاتها

إعداد

سندس محمد سعيد الحلبي

إشراف

الدكتور محمد مازن المحاييري

2023-2022م

الجمهورية العربية السورية
جامعة دمشق
كلية الهندسة الميكانيكية والكهربائية
قسم هندسة الحواسيب والأتمتة

التلخيص الآلي للنصوص العربية باستخدام الأنطولوجيا

أطروحة مقدمة لنيل درجة الدكتوراة في هندسة الحواسيب والأتمتة

اختصاص هندسة الحواسيب وشبكاتها

إعداد الطالبة : سنده محمد سعيد الحلبي

لجنة الحكم:

- | | | |
|------------------------|--|--------------|
| د. سمير كرمان | الأستاذ في قسم هندسة الحواسيب والأتمتة - كلية الهندسة الميكانيكية والكهربائية - جامعة دمشق | عضواً |
| د. محمد مازن المحاييري | الأستاذ المساعد في قسم هندسة الحواسيب والأتمتة - كلية الهندسة الميكانيكية والكهربائية - جامعة دمشق | عضواً مشرفاً |
| د. بسام لالا | الأستاذ المساعد في قسم الهندسة الطبية - كلية الهندسة الميكانيكية والكهربائية - جامعة دمشق | عضواً |
| د. هيام خدام | المدرسة في قسم الحواسيب والأتمتة - كلية الهندسة الميكانيكية والكهربائية - جامعة دمشق | عضواً |
| د. رافة خازم | المدرسة في قسم الحواسيب والأتمتة - كلية الهندسة الميكانيكية والكهربائية - جامعة دمشق | عضواً |

كلمة شكر

في نهاية هذا العمل لا بدّ لي من أتوجه أولاً بالحمد والشكر لله عزّ وجلّ على توفيقه لي لأصل إلى ماوصلت له.

كما أتوجه بالشكر والامتنان إلى المشرفين: الأستاذ الدكتور محمد مازن المحاييري حفظه الله ورعاه، والأستاذ الدكتور فيصل العباس رحمه الله، فقد كان لإشرافهم وتوجيهاتهم الدور الأساسي في إنجاز هذا البحث ونجاحه. وأشكر الأساتذة أعضاء لجنة التحكيم على تفضلهم بقبول تحكيم البحث. وإثراء هذا البحث بملاحظاتهم القيمة.

كما أتقدم بالشكر الجزيل إلى إدارة كلية الهندسة الميكانيكية والكهربائية متمثلةً: بالسيد العميد، والسيدان الوكيل العلمي والوكيل الإداري، ورئيس قسم هندسة الحواسيب، على تسهيل هذا العمل ونجاحه.

والشكر الكبير إلى أبي العظيم الدكتور محمد سعيد الحلبي قدوتي وسندي وأمي رمز العطاء والحنان، في وصفهما يتوه اللسان، أكرمهما ياربي بتمام فضلك فعمري لا يكافئهما لوأوتيت من العمر ثان.

وأتقدم بخالص شكري وامتناني إلى زوجي العزيز على صبره ومساندته وتشجيعه ودعمه لي، وابنتي وقرة عيني أمل حياتي، احفظهما ياربي فهما الأمان والأمل.

وأشكر شقائق روحي إخوتي وأخواتي يارب احفظهم واحفظ أبناءهم وبناتهم وأزواجهم وزوجاتهم وجميع أفراد عائلتي بحفظك يامولاي.

والشكر موصول إلى جميع أصدقائي وزملائي وأساتذتي، وأسرة هيئة الموسوعة العربية، وكل من ساندني وقدم لي العون في تنفيذ هذا العمل.

فهرس المحتويات

8	قائمة الجداول
9	قائمة الأشكال
10	قائمة المعادلات
12	قائمة المصطلحات
14	الملخص
15	الباب الأول : الدراسة المرجعية
16	الفصل الأول: مقدمة البحث
16	1-1 مقدمة
18	2-1 ضرورة البحث
18	3-1 الأهداف
19	4-1 المساهمات
20	5-1 منهجية البحث
21	6-1 محتويات الأطروحة
24	الفصل الثاني: الخلفية النظرية
24	1-2 مقدمة
24	2-2 التلخيص الآلي
25	3-2 تصنيف التلخيص الآلي
27	4-2 أنظمة التلخيص الآلي
28	5-2 أهم خصائص التلخيص الآلي
30	6-2 تقنيات تلخيص النصوص
31	1-6-2 التقنيات الإحصائية
32	2-6-2 تقنيات التعلم الآلي
34	3-6-2 التقنيات الدلالية
35	4-6-2 الذكاء الجماعي
36	7-2 تقييم ملخصات النصوص:
37	1-7-2 التقييم الجوهري
40	2-7-2 التقييم الخارجي
41	الفصل الثالث: الدراسات السابقة
41	1-3 مقدمة
41	2-3 أهم مشاريع التلخيص الآلي للنصوص ودراساتها

42	1-2-3 التقنيات الإحصائية
44	2-2-3 تقنيات التعلم الآلي
45	3-2-3 التقنيات الدلالية
47	4-2-3 اعتماد الذكاء الجماعي
47	3-3 الأبحاث والتقنيات المستخدمة لتلخيص النصوص العربية
47	1-3-3 التقنيات الإحصائية
47	2-3-3 تقنيات التعلم الآلي
48	3-3-3 التقنيات الدلالية
49	4-3-3 تقنيات الذكاء الجماعي
49	4-3 أهم أدوات وأنظمة التلخيص الآلي
53	3-5 وصف المشكلة وحلها
55	الباب الثاني: توليد ملخص آلي للنصوص العربية
56	الفصل الرابع: نموذج عام للملخص الآلي
56	1-4 مقدمة
56	2-4 نموذج التلخيص الآلي للنصوص العربية
57	1-2-4 توليد أنطولوجيا عربية
58	2-2-4 بناء نظام تلخيص آلي للنصوص العربية
59	3-2-4 تقييم أنظمة التلخيص المقترحة
60	4-2-4 تطبيق نظام التلخيص على نصوص الموسوعة العربية
61	الفصل الخامس: توليد أنطولوجيا عربية
61	1-5 مقدمة
62	2-5 تعريف الأنطولوجيا
63	3-5 تعلم الأنطولوجيا
64	4-5 شبكة الكلمات العربية
65	5-5 الموسوعة العربية
66	6-5 الأبحاث والدراسات السابقة لتوليد أنطولوجيا عربية
67	7-5 توليد الأنطولوجيا العربية بالتقابل بين شبكة الكلمات العربية والموسوعة العربية
68	1-7-5 توليد خصائص الموسوعة العربية
71	2-7-5 تجهيز بيانات شبكة الكلمات العربية
71	3-7-5 تعلم الأنطولوجيا بالتقابل بين خصائص الموسوعة وشبكة الكلمات
73	4-7-5 تقييم الأنطولوجيا العربية
74	الفصل السادس: توليد الملخص الآلي للنصوص العربية اعتماداً على الأنطولوجيا العربية

1-6	مقدمة	74
2-6	مراحل النظام	74
3-6	المعالجة الأولية للوثيقة	75
1-3-6	التقطيع	75
2-3-6	التقييس	75
3-3-6	إزالة كلمات التوقف	76
4-3-6	التجريد	76
4-6	التحليل واستخراج الخصائص وصياغتها	76
1-4-6	خاصية العبارات المفتاحية	78
2-4-6	خاصية التشابه مع العنوان	80
3-4-6	خاصية موقع الجملة	82
4-4-6	خاصية طول الجملة	82
5-4-6	خاصية العبارات البارزة	83
6-4-6	خاصية العبارات غير المهمة	84
5-6	تقييم الجمل وتوليد الملخص	84
الفصل السابع:	تحليل النتائج وتقييمها	86
1-7	مقدمة	86
2-7	تقييم التلخيص الآلي للنصوص العربية وتحليل نتائجه	86
1-2-7	مجموعة البيانات	87
2-2-7	التجارب والنتائج	88
3-2-7	تحليل النتائج وتقييمها	90
3-7	تقييم الأنطولوجيا العربية المتولدة وتحليل نتائجها	90
1-3-7	مجموعة البيانات	91
2-3-7	نتائج الأنطولوجيا العربية المتولدة	91
4-7	تطبيق التلخيص الآلي على الموسوعة العربية	92
الفصل الثامن:	الخلاصة والخاتمة	93
1-8	مقدمة	93
2-8	الخلاصة	93
3-8	الخاتمة:	95
4-8	الآفاق المستقبلية:	95
المراجع والمصادر		96
ملاحق البحث		105

الملحق (A): قائمة الجداول.....	106
الملحق (B): الإجراءات الأساسية	112

قائمة الجداول

List Of Tables

- الجدول (1-9) : أهم مشاريع ودراسات التلخيص الآلي للنصوص.....106
- الجدول (2-9) : أهم دراسات ومشاريع تلخيص النصوص العربية.....108
- الجدول (3-9) جمل الملخص الناتجة عن تلخيص مستند العينة باعتماد الأنطولوجيا109
مع قيم أوزانها ونسب خصائصها.
- الجدول (1-7): مقارنة مقاييس الدقة والاستدكار والمعامل F لأنظمة التلخيص الثلاثة91
(التلخيص باعتماد الأنطولوجيا، والتلخيص الإحصائي، وتلخيص نظام لخصلي).
- الجدول (4-9): عينات من مجموعات المفاهيم والمصطلحات المرتبطة في الأنطولوجيا العربية...110
- الجدول (5-9): عينات من علاقات التشابه والترافق في الأنطولوجيا العربية المتولدة.....111

قائمة الأشكال

List Of Figures

- الشكل (1-2) أنواع التلخيص 27
 - الشكل (2-2) بنية أنظمة التلخيص 27
 - الشكل (3-2) تقنيات تلخيص النصوص 31
 - الشكل (1-4) مراحل نموذج التلخيص الآلي للنصوص العربية 57
 - الشكل (2-4) مراحل توليد الأنطولوجيا العربية 58
 - الشكل (3-4) مراحل التلخيص الآلي 59
 - الشكل (1-5) خطوات المعالجة الأولية للنصوص العربية 68
 - الشكل (2-5) مراحل تجهيز بيانات شبكة الكلمات العربية 71
 - الشكل (3-5) خوارزمية تعلم الأنطولوجيا 72
 - الشكل (1-6) خطوات المعالجة الأولية في تلخيص النصوص 75
 - الشكل (2-6) خوارزمية إعادة مصطلحات الجمل إلى مفاهيمها 78
 - الشكل (1-7) عينة النص الأصلي 88
 - الشكل (2-7) الملخص باستخدام الأنطولوجيا 89
 - الشكل (3-7) الملخص باستخدام التلخيص الإحصائي 89
 - الشكل (4-7) الملخص باستخدام لخصلي Lakhasly 89
 - الشكل (5-7) مقارنة نتائج تقييم التلخيص باعتماد الأنطولوجيا مع التلخيص الإحصائي وتلخيص.. 91
- لخصلي.

قائمة المعادلات

LIST OF EQUATIONS

- المعادلة (1-2): علاقة حساب الدقة.....37
- المعادلة (2-2): علاقة حساب الاستدكار.....38
- المعادلة (3-2): علاقة حساب المعامل F.....38
- المعادلة (4-2): علاقة حساب تشابه التجيب.....39
- المعادلة (5-2) : علاقة حساب التداخل بين وحدتين.....39
- المعادلة (6-2): علاقة حساب ROUGE-N.....39
- المعادلة (1-5): علاقة حساب تثقيل المصطلحات في المستندات69
- المعادلة (2-5) :علاقة حساب الترافق بين مصطلحين.....70
- المعادلة (3-5) علاقة حساب التشابه بين مصطلحين.....71
- المعادلة (1-6): علاقة حساب تثقيل المصطلحات في الجمل79
- المعادلة (2-6): علاقة حساب تثقيل المفاهيم في الجمل80
- المعادلة (3-6) : علاقة حساب تشابه الجملة مع عنوان المستند اعتماداً على شعاعي مصطلحاتهما.....80
- المعادلة (4-6): علاقة حساب تشابه شعاعي المصطلحات.....81
- المعادلة (5-6): علاقة حساب تشابه الجملة مع عنوان المستند اعتماداً على شعاعي مفاهيمها.....81
- المعادلة (6-6): علاقة حساب تشابه شعاعي المفاهيم.....82
- المعادلة (7-6): علاقة حساب خاصية طول الجملة اعتماداً على عدد المصطلحات.....83
- المعادلة (8-6): علاقة حساب خاصية طول الجملة اعتماداً على عدد المفاهيم.....83
- المعادلة (9-6): علاقة حساب خاصية العبارات البارزة83
- المعادلة (10-6) : علاقة حساب خاصية العبارات غير المهمة.....84

- المعادلة (6-11): علاقة تثقيل الجمل اعتماداً على خصائصها.....84
- المعادلة (7-1): علاقة حساب الدقة لتقييم التلخيص الآلي.....86
- المعادلة (7-2): علاقة حساب الاستدكار لتقييم التلخيص الآلي.....87
- المعادلة (7-3) : علاقة حساب المعامل F لتقييم التلخيص الآلي.....87

قائمة المصطلحات

List Of Terms

Abstractive	تجريدي
Analyzer	المحلل
Arabic natural language processing (ANLP)	معالجة اللغة الطبيعية للعربية
Arabic Wordnet	شبكة الكلمات العربية
Automatic Summarization	التلخيص الآلي
bacterial foraging optimization	البحث الأمثل عن البكتيريا
Chemotaxis	الإنجذاب الكيميائي
Clustering	التجميع
Discourse theory	نظرية الخطاب
Document Understanding Conferences (DUC)	مؤتمرات فهم المستند
Elimination-dispersal	إلغاء التشتت
Essex Arabic Summaries Corpus (EASC)	مجموعة الملخصات العربية من إسكس
Extractive	استخراجي
Extrinsic Evaluation	التقييم الخارجي
Feature extraction	استخراج الخصائص
Indicative	الموجز
Informative	المعلوماتي
Intrinsic Evaluation	التقييم الجوهري
Key-phrases	العبارات المفتاحية
Latent Semantic Analysis (LSA)	التحليل الدلالي الكامن
Lexical Chain	السلسلة المعجمية
Normalization	التقييس
Ontology	أنطولوجيا

optimization algorithm	خوارزمية تحسين الوقواق
Parser	المفسر
particle swarm optimization	تحسين حشد الجسيمات
Part-of-Speech (POS)	نوع الكلام
post processing	المعالجة اللاحقة
Precision	الدقة
Preprocessing	المعالجة الأولية
Recall	الاستدكار
Recall-Oriented Understudy for Gisting Evaluation (ROUGE)	التقييم الموجه بالاستدكار
Reproduction	الاستنساخ
Rhetorical structure theory (RST)	نظرية البنية البلاغية
Segmented Discourse Representation Theory (SDRT)	نظرية البنية الخطابية المجزأة
Stemming	التجريد
StopWord Removal	إلغاء كلمات التوقف
Summarization	التلخيص
Swarm intelligence	الذكاء الجماعي
Synthesizer	المولد
Tokenization	التقطيع
Vector Space Model (VSM)	النموذج الشعاعي
Wordnet	شبكة الكلمات

الملخص

تلخيص النصوص هو عملية استخلاص أو تجميع المعلومات المهمة من النصوص الأصلية وعرضها بشكل ملخص. يساعد التلخيص في الحصول على المعلومات المطلوبة بأقل وقت. يوجد الكثير من الأبحاث والدراسات والأنظمة الخاصة بالتلخيص الآلي للنصوص باللغات الأوروبية المتعددة وغيرها من اللغات العالمية، إلا أن الأبحاث والدراسات في مجال تلخيص النصوص العربية ضئيلة وتنمو ببطء (El-Haj et al., 2011a) واتبعت في معظمها الأسلوب الاستخراجي واستخدام تقنيات إحصائية وابتعدت عن الأسلوب التجريدي واستخدام التقنيات الدلالية استناداً إلى مصادر معرفية مثل الأنطولوجيا. لذا اقترح في هذا البحث نظام تلخيص آلي للنصوص العربية باعتماد الأنطولوجيا، فقد بُني نظام تلخيص استخراجي عام أحادي اللغة وأحادي المستند يستخدم تقنيات دلالية بالاعتماد على الأنطولوجيا، ووظفت فيه خصائص إحصائية ودلالية لتقليل الجمل وتقييم أهميتها استناداً إلى توارد مفاهيم المصطلحات الواردة فيها وتكرارها بدلاً من المصطلحات نفسها باستخدام علاقات دلالية بين المصطلحات تستخرج من الأنطولوجيا العربية وهي الترادف synonym والاشتمال holo_part والتضمين hyponym والعضو holo_member. وقد استخدمت أنطولوجيا الكلمات العربية (AWN) Arabic Wordnet المستخدمة على نحو واسع بعد توسيعها وإغنائها بمفاهيم وعلاقات مستنبطة من الموسوعة العربية باعتماد تقنية التقابل mapping. وقُيِّم النظام باستخدام مجموعة الملخصات العربية من Essex (EASC) (El-Haj et al., 2010) واعتماد مقاييس الدقة والاستدكار ومعامل F. وقد بينت النتائج أن متوسطات الدقة والاستدكار ومعامل F في التلخيص باعتماد الأنطولوجيا المقترح هي نسب جيدة مقارنة مع نسب أنظمة التلخيص الآلي للنصوص العربية المقارن معها، ومع أنظمة التلخيص عموماً.

الكلمات الرئيسية: التلخيص الآلي، التلخيص الآلي للنصوص العربية، التلخيص الاستخراجي، التلخيص التجريدي، التلخيص الدلالي، الأنطولوجيا، أنطولوجيا عربية، شبكة الكلمات العربية، التقابل، الموسوعة العربية.

الباب الأول

الدراسة المرجعية

الفصل الأول

مقدمة البحث

Introduction

1-1 مقدمة:

زاد انتشار البيانات على الإنترنت، فأدى ذلك إلى وجود كم هائل من المصادر المتاحة للمستخدمين ومتصفح الإنترنت، لذا أصبحت عملية التلخيص الآلي للنصوص ضرورية للتغلب على مشكلة تزايد المعلومات وتضخمها. يُعرّف الملخص عموماً على أنه: " النص المولد من نص أو مجموعة نصوص، والذي يتضمن المعلومات المهمة من النص الأصلي، وهو بطول لا يزيد على نصف طول النص الأصلي وعادة يكون أقل من ذلك " (Radev et al., 2002). ويمكن أن تتضمن الملخصات أي نوع آخر من الوسائط مثل الصوت والفيديو.

تُصنّف أنظمة التلخيص حسب خصائص عدة هي: الدخل، الخرج، العمومية، اللغة، الطريقة. هذا يسمح بتوصيف الملخصات بمجموعة واسعة من الخصائص (Lloret et al., 2008). مثلاً يمكن أن يكون التلخيص أحادي المستند أو متعدد المستندات، وبحسب العمومية يمكن أن يُصنّف إلى ملخص عام وملخص مقاد بالاستعلام. بخصوص شكل الخرج يتم التمييز بين الملخصات الدلالية indicative والمعلوماتية informative. علاوة على ذلك يمكن التحدث عن تلخيص أحادي اللغة وتلخيص متعدد اللغات. كما يمكن تصنيف طرائق التلخيص إلى استخراجية وتجريدية. تتضمن طرائق التلخيص الاستخراجي اختيار الجمل المهمة من المستند المصدر اعتماداً على خصائص إحصائية ولغوية ودمجها إلى شكل موجز (Gupta et al., 2010). بينما تختلف طريقة التلخيص التجريدي عن الاستخراجي بتوفير ملخصات لها درجة من الاستدلال حول المعرفة الخلفية الموجودة في المستند الأصلي (Bawakid, 2011)، تُستخدم فيه طرائق لغوية لفحص النص وإيجاد مفاهيم وتعابير جديدة لوصفه على نحو أفضل.

طورت تقنيات عدة لتلخيص النصوص من ضمنها التقنيات الدلالية باعتماد الأنطولوجيا. قدم هذا البحث نموذجاً لتلخيص النصوص العربية تضمن ثلاث مراحل هي: توليد أنطولوجيا عربية. وبناء نظام لتلخيص النصوص العربية بطريقة دلالية اعتماداً على الأنطولوجيا العربية المولدة و تقييم نتائج تلخيص النظام وتطبيقه على نصوص الموسوعة العربية لتلخيصها.

اقترحت طريقة لتوليد أنطولوجيا عربية بالتقابل بين شبكة الكلمات العربية Arabic Wordnet وخصائص مقالات الموسوعة العربية، كما وُلدَ توليد نظام تلخيص استخراجي عام أحادي اللغة وأحادي المستند يستخدم تقنيات دلالية بالاعتماد على الأنطولوجيا العربية، موظفاً خصائص إحصائية لتثقيل الجمل وتقييم أهميتها استناداً إلى توارد مفاهيم المصطلحات الواردة فيها وتكرارها بدلاً من المصطلحات نفسها باستخدام علاقات دلالية بين المصطلحات تستنبط من الأنطولوجيا العربية المقترحة وهي: الترادف، والاشتغال، والتضمين، والعضو. وبذلك يتم الدمج بين الطريقة الإحصائية والدلالية في التلخيص الآلي من أجل تحسين جودة التلخيص.

وُقِّمَ النظام باستخدام مجموعة الملخصات العربية من (Essex (EASC (El-Haj et al., 2010 واعتماد مقاييس الدقة والاستدكار ومعامل F. وقد بينت النتائج أن متوسطات الدقة والاستدكار ومعامل F في التلخيص باعتماد الأنطولوجيا المقترحة هي نسب جيدة مقارنة مع نسب أنظمة التلخيص الآلي للنصوص العربية المقارن معها.

1-2 ضرورة البحث:

اللغة العربية هي اللغة الأم لأكثر من 300 مليون شخص حول العالم وهي واحدة من اللغات الست الرسمية لدى الأمم المتحدة (Farghaly et al., 2009). يستدعي التعامل معها تحديات كثيرة في مجال معالجة اللغات الطبيعية (Farghaly et al., 2009)، ويتطلب أعمال معالجة مختلفة عند تنفيذ عمليات مثل: الترجمة الآلية والتلخيص الآلي للنصوص العربية.

كما بينت الدراسات قلة الأبحاث والمشاريع التي اهتمت بتلخيص النصوص العربية وحدثتها مقارنة مع اللغات الأخرى، وابتعت في معظمها الأسلوب الاستخراجي واستخدام التقنيات الإحصائية، وابتعدت عن الأسلوب التجريدي واستخدام التقنيات الدلالية والاعتماد على مصادر معرفية كالأنطولوجيا، وذلك بسبب عدم وجود معايير موحدة للغة العربية، وبسبب ضعف الأدوات المطبقة في معالجة اللغات الطبيعية للغة العربية.

لذا كان لابد من الاستفادة من دراسات التلخيص للنصوص باللغات العالمية الأخرى عموماً وباللغة العربية خصوصاً. وإيجاد طريقة تلخيص آلي للنصوص العربية تراعى فيها خصوصية اللغة العربية وتستخدم تقنيات دلالية باستخدام مصادر معرفية كالأنطولوجيا العربية، للحصول على غنى معرفي من خلال العلاقات الدلالية بين المصطلحات، التي تساهم في تحسين تقييم الجمل المهمة من النصوص العربية على أساس مفاهيم المصطلحات وعلاقاتها الدلالية. من أجل الحصول على ملخصات ذات جودة أفضل.

1-3 الأهداف:

يهدف هذا البحث لإيجاد نموذج لتطوير ملخص آلي للنصوص العربية باستخدام أنطولوجيا عربية لتحسين جودة التلخيص باستخدام العلاقات الدلالية بين المفاهيم العربية، وتوليد هذا الملخص اعتماداً على النموذج المقترح، ومن ثم تجريب نتائج التلخيص واختبارها وتقييم الملخص وتطبيقه على الموسوعة العربية لتلخيصها.

لذا تتضمن أهداف هذا البحث مايلي:

1- تحليل الدراسات السابقة لمعرفة الوضع الحالي في المواضيع المرتبطة ببناء ملخص آلي للنصوص العربية.

2- إيجاد نموذج عام لتوليد الملخص بالاعتماد على الدراسات والتجارب السابقة في توليد ملخصات آلية باللغات المختلفة مع مراعاة خصائص اللغة العربية.

3- توليد أنطولوجيا عربية كمصدر معرفي بتوسيع شبكة الكلمات العربية Arabic WordNet بالتقابل بين مفاهيمها وخصائص مقالات الموسوعة العربية.

4- توليد ملخص نصوص عربية اعتماداً على النموذج المقترح واستخدام الأنطولوجيا العربية كمصدر معرفي لتحسين أداء الملخص وجودة الملخصات من خلال العلاقات الدلالية بين المصطلحات والعبارات والجمل اعتماداً على العلاقات بين المفاهيم العربية في الأنطولوجيا العربية.

5- اختبار الملخص وتجريبه وتقييمه

6- تطبيق نظام التلخيص على نصوص الموسوعة العربية لتلخيصها.

1-4 المساهمات:

يساعد التلخيص الآلي للنصوص على توفير وقت القارئ وجهده، ومن ثمَّ يمكنه من اتخاذ القرارات بشكل أسرع. لذا أصبح هناك اهتمام بالتلخيص الآلي للنصوص بشكل عام وللنصوص العربية بشكل خاص. يهدف البحث إلى التركيز على التلخيص الآلي للنصوص العربية ببناء نظام تلخيص آلي دلالي باعتماد الأنطولوجيا العربية. وكان لهذا النظام إسهامات نلخصها فيما يلي:

1- إن توليد ملخص آلي للنصوص العربية مهم لتحسين المحتوى الرقمي العربي على الإنترنت ولخدمة القارئ العربي واللغة العربية، وذلك من خلال استخدامه لإيجاد ملخصات للنصوص العربية متناسقة وذات جودة مناسبة. وهو يفيد في تحسين عمليات استرجاع المعلومات أثناء البحث عن المعلومات حيث يوفر الوقت والجهد في الحصول على المعلومات المهمة.

2- إن اعتماد تقنية التقابل في توليد أنطولوجيا عربية كمصدر معرفي يعد إسهاماً مهماً في مجال تعلم الأنطولوجيا العربية، فقد طبقت هذه التقنية بتوسيع شبكة الكلمات العربية Arabic WordNet التي تعد

أشهر أنطولوجيا عربية مترجمة وعامة وتستخدم على نحو واسع، بإضافة مفاهيم ومصطلحات عربية مستخرجة من خصائص مقالات الموسوعة العربية وعلاقات دلالية. كما إن استخدام الموسوعة العربية التي تعدّ أهم مصدر معلومات عربي غني وشامل لمجالات المعارف الإنسانية كلها في تقنية التقابل، وليس استخدام موسوعات مترجمة مثل الويكيبيديا كمصدر للمعلومات كما اتبع في الدراسات المشابهة السابقة يعد إسهاماً مهماً لإغناء الشبكة بمفاهيم عربية ومصطلحات غير موجودة فيها.

3- إن توليد أنطولوجيا عربية بحد ذاته عمل مهم ويسهم في دعم التطبيقات التي تعتمد على معالجة اللغة العربية مثل الوب الدلالي والترجمة والتلخيص الآلي.

4- الدمج بين الطريقة الإحصائية والدلالية في التلخيص الآلي للنصوص العربية له دور في تحسين جودة التلخيص مقارنة مع أنظمة التلخيص العربية، وذلك من خلال توليد ملخص آلي للنصوص العربية يعتمد الأنطولوجيا موظفاً خصائص إحصائية لتثقيل الجمل وتقييم أهميتها استناداً إلى توارد مفاهيم المصطلحات الواردة فيها وتكرارها بدلاً من المصطلحات نفسها باستخدام علاقات دلالية بين المصطلحات تستنبط من الأنطولوجيا العربية المقترحة وهي الترادف والاشتغال والتضمين والعضو.

5- أسهم هذا البحث في اعتماد الأنطولوجيا في التلخيص العام من خلال تمثيل جمل المستند بمفاهيمها، وليس التلخيص باعتماد الاستعلام كما هو مطبق في أنظمة التلخيص العربية التي تستخدم الأنطولوجيا لتوسيع الاستعلام باستخدام المرادفات كما هو موضح في الدراسة المرجعية لأنظمة التلخيص العربية.

1-5 منهجية البحث:

سوف يبدأ هذا البحث بتحليل الدراسات السابقة لمعرفة الوضع الحالي في المواضيع المرتبطة ببناء ملخص آلي للنصوص العربية، وبعدها يقترح نموذجاً عاماً لتوليد الملخص الآلي للنصوص العربية باعتماد الأنطولوجيا، مستفيداً من الدراسات والتجارب السابقة في توليد ملخصات آلية باللغات المختلفة مع مراعاة خصائص اللغة العربية.

تولد بعد ذلك أنطولوجيا عربية مناسبة لاستخدام علاقاتها الدلالية بين مفاهيم اللغة العربية في تحسين عملية التلخيص بتوسيع شبكة الكلمات العربية Arabic WordNet بالتقابل بين مفاهيمها وخصائص مقالات الموسوعة العربية.

يطبق بعدها النموذج العام المقترح لتوليد ملخص آلي استخراجي عام أحادي اللغة وأحادي المستند مستخدماً الأنطولوجيا العربية بوصفها مصدراً معرفياً. وحساب خصائص الجمل لتثقيل الجمل وتقييم أهميتها لتضمينها في الملخص استناداً إلى تواردها مفاهيم المصطلحات الواردة فيها وتكرارها باستخدام العلاقات الدلالية استناداً إلى الأنطولوجيا العربية. كما يتم توليد نظام تلخيص إحصائي للنصوص العربية من أجل التقييم والمقارنة، يعتمد هذا النظام الطريقة الإحصائية لإيجاد ملخص استخراجي عام أحادي المستند وأحادي اللغة، تحسب خصائص الجمل فيه استناداً إلى تواردها وتكرار المصطلحات والعبارات الواردة فيها.

وأخيراً تقييم الملخص الآلي باستخدام مجموعة الملخصات العربية من (Essex (EASC El-Haj et al., 2010). باستخدام مقاييس الدقة والاستنكار ومعامل F، ومقارنة نتائج مقاييس التقييم مع نتائج التلخيص الإحصائي المقترح ونظام التلخيص (الخصلي) المشهور. وتطبيق التلخيص الآلي على مقالات من الموسوعة العربية لتلخيصها.

1-6 محتويات الأطروحة:

تُقسم هيكلية الأطروحة إلى جزئين أساسيين، يتضمن الجزء الأول مقدمة وخلفية نظرية ومراجعة الأدبيات والأعمال والدراسات السابقة، ويتناول الجزء الثاني الإسهامات الأساسية في توليد ملخص آلي للنصوص العربية والذي يتضمن إنشاء نموذج الملخص باعتماد الأنطولوجيا، وتوليد أنطولوجيا عربية بتوسيع شبكة الكلمات العربية Arabic Wordnet وخصائص نصوص الموسوعة العربية، وكيفية تنفيذ نموذج التلخيص الآلي للنصوص العربية المقترح باعتماد الأنطولوجيا العربية المولدة، والنتائج والتقييم.

تم هيكلة الأطروحة في أبواب وفصول كما يلي:

الباب الأول: الدراسة المرجعية

الفصل الأول: المقدمة Introduction : يتناول هذا الفصل مقدمة عامة عن البحث ، ويبين ضرورة

البحث وأهدافه ، ويسلط الضوء على المساهمات الناتجة عنه.

الفصل الثاني: الخلفية النظرية Background: يعرض هذا الفصل أنواع التلخيص الآلي وتصنيفاته الأساسية

وأهم التقنيات والخصائص المستخدمة في عمليات التلخيص الآلي للنصوص.

الفصل الثالث: الدراسات المرتبطة بالبحث Related Works: ويقدم هذا الفصل عرضاً لأهم المشاريع

ودراسات التلخيص الآلي للنصوص وأدواتها.

الباب الثاني: توليد ملخص آلي للنصوص العربية

الفصل الرابع: نموذج عام للملخص الآلي General Model for Automatic Summarizer: يقدم في هذا

الفصل نموذجاً لتلخيص النصوص العربية اعتماداً على الطريقة الدلالية باستخدام الأنطولوجيا

الفصل الخامس: توليد أنطولوجيا عربية Generation of Arabic Ontology: يقدم هذا الفصل طريقة

لتوليد الأنطولوجيا العربية اعتماداً على التقابل بين شبكة الكلمات العربية وخصائص الموسوعة

العربية.

الفصل السادس: توليد الملخص الآلي العربي اعتماداً على الأنطولوجيا العربية Generation of

Automatic Arabic Summarizer Using Arabic Ontology: يتضمن هذا الفصل طريقة بناء

نظام التلخيص الآلي للنصوص العربية وكيفية استخدام الأنطولوجيا العربية المتولدة فيه.

الفصل السابع: تحليل النتائج وتقييمها Resultes Analyses and Evaluation: يقدم هذا الفصل تحليلاً

لنتائج التلخيص الآلي للنصوص العربية وتقييمها، كما يتناول شرحاً للأنطولوجيا العربية المتولدة

الفصل الثامن: الخلاصة والخاتمة Summary and Conclusion: نقدم هذا الفصل خلاصة مأنجز في هذا

البحث والخاتمة والآفاق المستقبلية.

الفصل الثاني

الخلفية النظرية

Background

1-2 مقدمة:

يقدم هذا الفصل شرحاً مبسطاً عن التلخيص الآلي للنصوص وأنواعه وخطواته، وتصنيفاته الأساسية، كما يقدم أهم التقنيات والخصائص المستخدمة في عمليات التلخيص الآلي للنصوص، ويوضح طرائق تقييم الملخصات الآلية.

2-2 التلخيص الآلي:

أصبحت الملخصات شائعة الاستخدام ولها وجود قوي في هذه الأيام وكأمثلة عليها: إيجاز المقالات، العناوين الرئيسية لأخبار التلفزيون، ومراجعات الكتب، وأدلة التسوق. يُعرّف الملخص عموماً على أنه: " النص المولد من نص أو مجموعة نصوص، والذي يتضمن المعلومات المهمة من النص الأصلي، وهو بطول لا يزيد على نصف طول النص الأصلي وعادة يكون أقل من ذلك (Radev et al., 2002)". ويمكن أن تتضمن الملخصات أي نوع آخر من الوسائط مثل الصوت والفيديو.

مع السرعة الكبيرة لانتشار البيانات على الإنترنت، أصبح هناك كم هائل من المصادر المتاحة للمستخدمين ومتصفحي الإنترنت. لذا أصبحت عملية التلخيص الآلي للنصوص ضرورية للتغلب على مشكلة تزايد وتضخم المعلومات. تختصر عملية التلخيص الوقت والجهد المطلوبين لتحديد الأجزاء البارزة والأكثر أهمية من النصوص.

2-3 تصنيف التلخيص الآلي:

لا يوجد تصنيف واحد معين لأنواع التلخيص، يختلف تصنيف أنواع التلخيص حسب وجهة النظر (Nazari et al., 2019). وتتغير مهام التلخيص الآلي باعتماد محددات عدة هي: الدخل، والخرج، والعمومية، واللغة، والتفصيل، والمعارف. وسوف نناقش هذه الأنواع فيما يلي:

الدخل:

يمكن أن يكون التلخيص أحادي المستند أو متعدد المستندات (Kumar et al., 2016). وقد كانت الاهتمامات السابقة معتمدة على تلخيص أحادي المستند، تولد الملخصات في مثل هذه الأنظمة من مستند مصدر وحيد. بينما قام المطورون مؤخراً بتطوير ملخصات تعتمد مستندات متعددة المصادر. تؤخذ في التلخيص متعدد المستندات مستندات مختلفة تشترك في الموضوع نفسه كدخل للنظام. وهذا يزيد من تعقيد النظام وذلك لأنه يتطلب إلغاء التكرار عبر المستندات وإعادة صياغة المحتويات للحصول على ملخص متماسك.

التفصيل:

يمكن أن يكون التلخيص دلاليًا indicative أو معلوماتيًا informative، تقدم أنظمة التلخيص الدلالية فقط الفكرة المهمة من النص، بينما تنتج الأنظمة المعلوماتية نظرة عامة عن المواضيع التي يغطيها النص. يتراوح الطول المثالي في نوع التلخيص الدلالي بين 5 و 10% من الطول الأصلي للمستند (Jezek et al., 2008)، بينما يكون طول الملخص المعلوماتي حوالي 20 إلى 30% من طول المستند الأصلي (Fan et al., 2005).

الخرج :

يُصنّف التلخيص باعتماد الخرج الناتج إلى: ملخص استخراجي Extractive أو تجريدي Abstractive (Gupta et al., 2010). تختار الجمل المهمة في طرائق التلخيص الاستخراجي من المستند المصدر اعتماداً على خصائص إحصائية ولغوية ودمجها على شكل ملخص، بينما توفر طريقة التلخيص التجريدي ملخصات لها درجة

من الاستدلال حول المعرفة الخلفية الموجودة في المستند الأصلي (Bawakid, 2011) وتستخدم طرائق لغوية لفحص النص وإيجاد مفاهيم وتعابير جديدة لوصفه على نحو أفضل.

العمومية:

يمكن أن تُصنف الملخصات إلى ملخصات عامة generic، أو ملخصات باعتماد الاستعلام Query-based. يُولد الملخص في التلخيص باعتماد الاستعلام باختيار الجمل التي تتوافق مع استعلام المستخدم (Gong et al., 2011)، فتحصل الجمل ذات الصلة بالاستعلام على فرصة أكبر لتضمينها في الملخص النهائي، ومن ثمّ لاتعطي أنظمة التلخيص باعتماد الاستعلام نظرة شاملة عن مفاهيم المستند، لأنها تركز على استعلام المستخدم. الهدف من التلخيص العام هو تلخيص النص كله دون ملاحظة استعلام أو موضوع معين (Jezek et al., 2008)، فليس لدى الملخصات العامة أي نظرة عن الموضوع وتتعامل مع المستند بوصفه وحدة نصية، ومن ثمّ يكون لكل المعلومات المستوى نفسه من الأهمية (Nenkova et al., 2011).

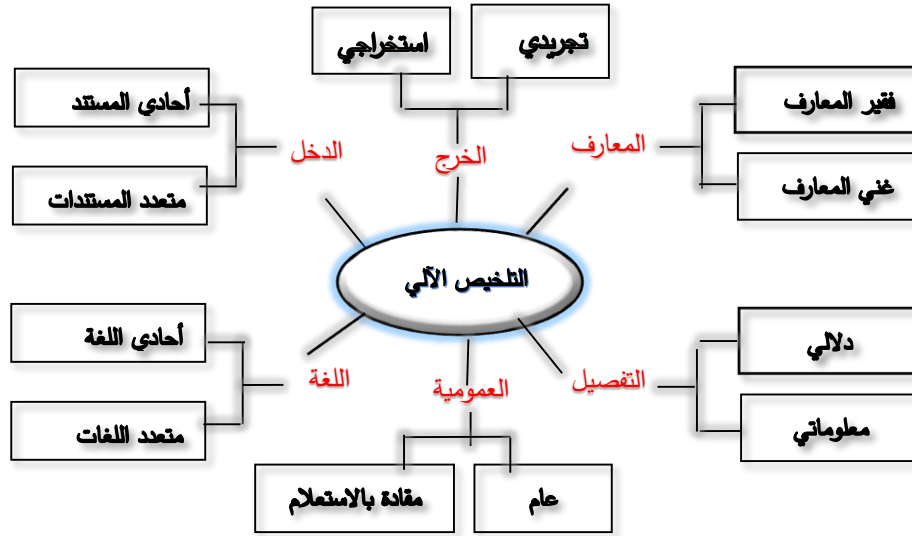
اللغة :

بالاعتماد على اللغة هناك ثلاثة أنواع من الملخصات هي: أحادية اللغة monolingual ومتعددة اللغات multi-lingual، وعبر اللغات cross-lingua (Jezek et al., 2008). في نظام التلخيص أحادي اللغة تكون لغة المستندات المصدر، والمستندات الهدف نفسها. في التلخيص متعدد اللغات يمكن أن تكون لغة المستند المصدر والملخص المولد في اللغة نفسها، ويستخدم النظام للتلخيص بعدد من اللغات. وفي الملخصات عبر اللغات مثل الملخصات متعددة اللغات إلا أنها تختلف عنها بأن لغة المصدر والهدف يجب أن تكون مختلفة.

المعارف:

يُصنّف التلخيص الآلي حسب استخدام التلخيص مصادر معرفية خارجية أم لا، فهناك تقنيات فقيرة معرفياً poor Knowledge techniques وهي التي لاتستخدم أي مصادر خارجية، بينما تُوظف التقنيات الغنية معرفياً rich knowledge techniques مصادر معرفية في عملية التلخيص مثل الموسوعات أو الأنطولوجيات التي تستخدم عادة للحصول على العلاقات الدلالية بين الكلمات، أو بين العبارات أو بين الجمل.

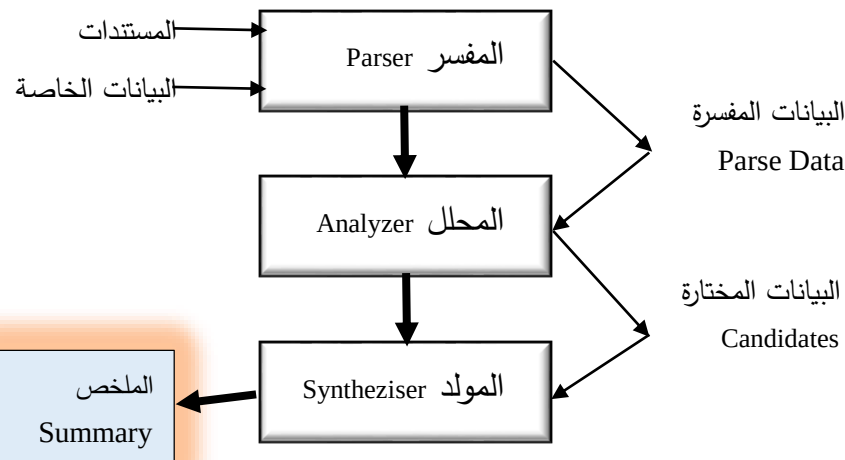
يبين الشكل رقم (1-2) أنواع أعمال التلخيص حسب المحددات المذكورة.



الشكل (1-2) أنواع التلخيص

4-2 أنظمة التلخيص الآلي:

تتشارك جميع أنظمة التلخيص الآلي في المراحل الثلاثة الأساسية للتلخيص وهي: إدخال المستندات والبيانات الخاصة بالعملية، والتحليل، وإخراج الملخصات. يبين الشكل رقم (2-2) أن هناك ثلاثة مكونات أساسية هي: المفسر Parser والمحلل Analyzer والمولد Synthesizer، يُشار عادة إلى أول خطوة بالمعالجة الأولية Preprocessing بينما يشار إلى الأخيرة بالمعالجة اللاحقة post processing.



الشكل (2-2) بنية أنظمة التلخيص

يغذى المفسر اللغوي أولاً بالمستندات المطلوب تلخيصها والبيانات الخاصة بالعملية مثل نسبة الضغط أو استعلامات المستخدم، ليقوم بمعالجتها وتجهيز المستندات بالشكل المناسب لمرحلة المحلل، كما تلغى البيانات غير الضرورية من المستندات والتي لاتعالج من قبل النظام في هذه المرحلة أيضاً. يطبق عادة المفسر اللغوي وحدة الكشف عن حدود الجملة على المستندات لتقسيمها إلى جمل، وبعد التقسيم إلى جمل، يمكن أن يشار إلى معلومات إضافية في كل جملة، مثل: ترتيب حدوثها ضمن النص، أو ضمن الفقرة، إذا وجد فيها أكثر من جملة. وإذا تطلب النظام بيانات إضافية كجزء من الدخل في مرحلة التحليل فعادةً ينجز التأثير ضمن وحدة المفسر اللغوي. وبعد تقطيع كل جملة إلى مصطلحات توضع تأشيريات، مثل: نوع الكلام (Part-of-Speech (POS.

تغذي البيانات المولدة المحلل Analyzer الذي يطبق الخوارزميات الأساسية للنظام مثل: تبسيط الجمل أو تقسيمها، أو ضغطها، وتقليل خصائص الجمل وترتيبها، كما توظف في هذه المرحلة مكونات إضافية في الحالات التي يستخدم فيها النظام مصادر معرفية خارجية للمساعدة في التلخيص، وتكون نتيجة المحلل كل ماتم ترشيحه وترتيبه من جمل أو عبارات أو مقاطع.

دور المولد synthesizer هو تنظيم الجمل والعبارات المرشحة والمرتببة وإظهارها بالشكل المناسب، مراعيًا الشروط المطلوب تحقيقها، مثل: نسبة الضغط، أو عدد الكلمات في الجمل.

2-5 أهم خصائص التلخيص الآلي

تستخلص أدوات وتقنيات تلخيص النصوص الجمل المفتاحية من النصوص المصدر، وتدمجها لتشكيل ملخصاً موجزاً. ويمكن بيان أهم الخصائص التي تُستخدم في اختيار الجمل المهمة بما يلي:

العبارات المفتاحية: العبارات المفتاحية هي قائمة مختصرة بالعبارات الأساسية المهمة التي توفر ملخصاً موجزاً للموضوع الرئيسي في المستند (Turney, 2000)، قد تكون كلمة واحدة أو مصطلحاً من كلمات عدة. إن إمكانية وجود فكرة أساسية في الجملة مشروطة باحتوائها على عبارة أو كلمات مفتاحية، وهذا يزيد من أهميتها بالنسبة للجمل الأخرى (El-Shishtawy et al., 2012). تعتمد درجة خاصية العبارة المفتاحية على عدد من

العوامل، بما في ذلك: تكرار العبارة، وعدد الكلمات فيها، وتكرار الكلمة ذات التوارد الأكبر فيها، وطول العبارة بالنسبة للجملة التي تحتويها (El-Shishtawy et al., 2012).

الموقع: تعتمد أهمية الجملة على موقعها في الفقرة/ المستند بغض النظر عن مجال/ موضوع المستند، اقترحت هذه الخاصية أولاً في (Baxendale, 1958) لتقييم الجمل. كما اعتمد موقع الجملة في مجال تلخيص النص الاستخراجي في الأنظمة: ((Abuobieda et al., 2012)، (Gupta et al., 2011)).

العبارات البارزة: يمكن أن يكون وجود بعض العبارات في الجملة (مثل: الأهم، بالتحديد، كملخص، كنتيجة، الأفضل) مؤشراً جيداً على أهمية المحتوى ((Lakshmi et al., 2015)، (Brazilay et al., 2015))، ومن ثم تُعطى الجمل التي تحتوي على هذه العبارات درجة أعلى مقارنةً بالجمل الأخرى.

التشابه مع العنوان: اقترحت هذه الخاصية أولاً في (Edmundson, 1969)، وعُرفت على أنها التشابه أوالتداخل بين جملة معينة وعنوان المستند ((Fattah et al., 2009)، (Abuobieda et al., 2012)، (Nobata et al., 2009)). تأتي أهمية هذه الخاصية من فكرة أنه إذا كانت الجملة تتكون من كلمات تظهر في العنوان فقد تكون جملة مهمة. علاوة على ذلك، إذا كانت الجملة تشترك في العبارة المفتاحية مع العنوان، فسيؤدي ذلك إلى زيادة أهميتها بشكل كبير.

طول الجملة: قد يؤثر طول الجملة على أهميتها. ومن ثم، قد تُستبعد الجمل الطويلة جداً أو القصيرة جداً من الملخص. فيمكن أن تزيد الجملة الطويلة جداً من محتواها المعلوماتي، ولكن يمكن أن يتجاوز طولها الحد المقبول عندما تصبح الجملة مفرطة في التفاصيل. أيضاً، تميل الجمل القصيرة إلى تضمين معلومات أقل مقارنةً بالجمل الأخرى، ومن ثم فهي أقل أهمية.

خاصية توارد المعلومات غير المهمة: بعض العبارات مثل (بالإضافة إلى ذلك، على سبيل المثال) هي علامات للكلام بمثابة كلمات شرح. تضعف مثل هذه العبارات الجملة لأنها توحي بأن الجملة القادمة هي معلومات تفصيلية إضافية ليست ضرورية في عملية التلخيص.

الأسماء: يمكن أن تُعدّ الجمل التي تحوي الكثير من الأسماء هي جمل مهمة تُضمّن في الملخص، مثل أسماء الأشخاص أو الأماكن وغيرها.

الحروف الكبيرة: يمكن أن تُعدّ الجمل التي تحوي كلمات بأحرف كبيرة مهمة في الملخص. هذه الخاصية لا تنطبق على اللغة العربية.

الخط: يمكن أن تُعدّ الجمل التي تحوي كلمات بنمط خط معين مثل: الغامق، المائل، تحته خط أوله حجم أكبر من قيمة معينة جملًا مهمة يمكن تضمينها في الملخص.

المعطيات الرقمية: يمكن أن تُعدّ الجمل التي تحتوي معطيات رقمية مثل 1،2،3.... ، و....iii، ii، i ، وألجمل التي تحتوي معلومات عن الأموال أو الربح أو الخسارة وغيرها مهمة وتُضمّن في الملخص.

مركزية الجملة: يحسب التشابه من أجل كل جملة مع الجمل الأخرى، وفي النهاية يحسب مجموع التشابهات لإيجاد قيمة الخاصية المركزية لهذه الجملة.

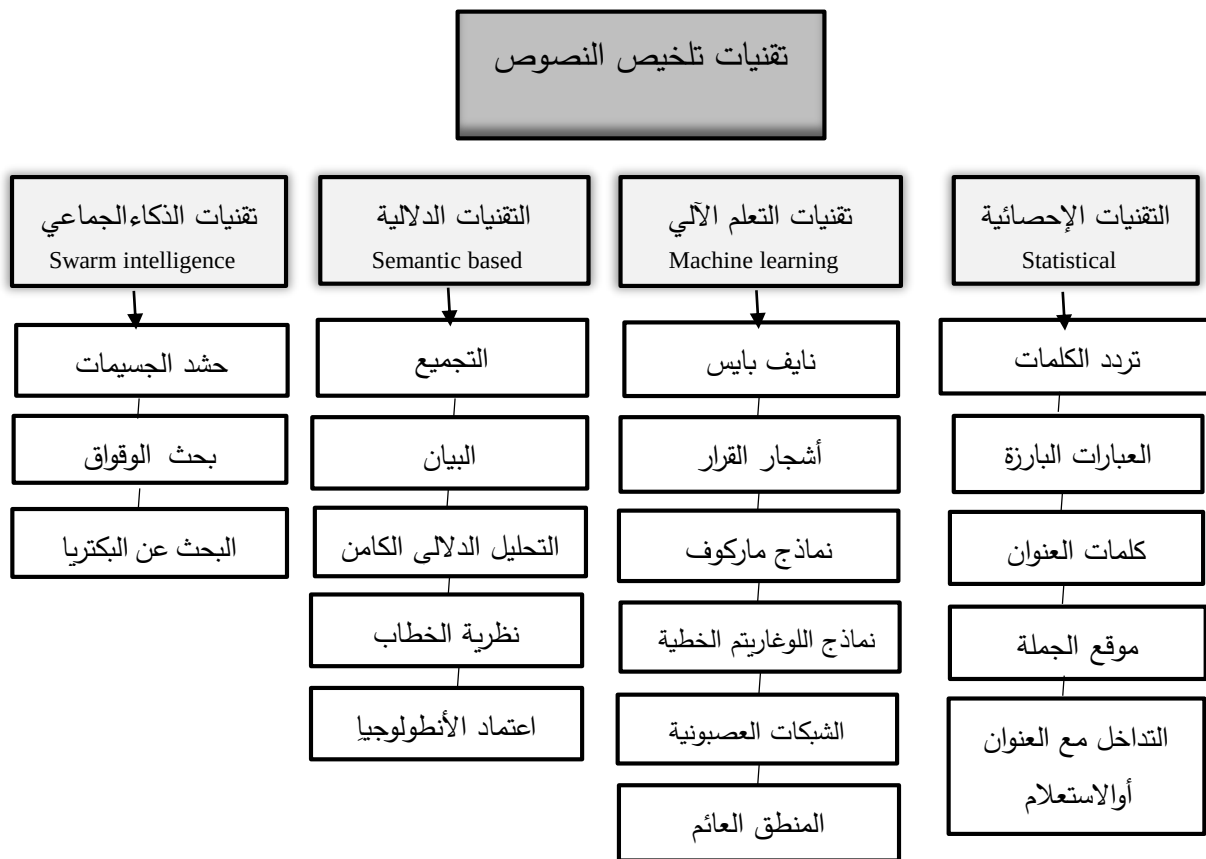
الضمير: يجب ألا تضمن الجمل التي تحتوي ضمائر مثل: هو هي هم وغيرها في الملخص دون ذكر الأسماء التي تعود عليها هذه الضمائر.

2-6 تقنيات تلخيص النصوص:

صنف الباحثون تقنيات التلخيص الآلي في (Nazari et al., 2019) إلى أربع فئات هي: إحصائية Statistical، وتعلم الآلة Machine Learning، واعتماد الدلالة Semantic-Based، واعتماد الذكاء الجماعي Swarm Intelligent-Based. يبين الشكل رقم (2-3) تقنيات تلخيص النصوص بشكل هرمي. وفيما يلي نبين شرحاً عن تقنيات التلخيص وأهم المشاريع في كل نوع.

2-6-1 التقنيات الإحصائية:

تتعامل هذه الطرائق مع بعض الخصائص الإحصائية Statistical methods للنص لتحديد الأجزاء المهمة في المستند. تهدف الطرائق الإحصائية لاختيار الجمل حسب خصائص فيزيائية وليس حسب معنى أو علاقة الكلمات والجمل، وهي استخراجية عموماً وتوظف بعض الخصائص مثل: تردد الكلمات، موقع الجمل، والكلمات البارزة وتداخلها مع عنوان المستند أو الاستعلام.



الشكل (2-3) تقنيات تلخيص النصوص

نبين فيما يلي بعض الخصائص الإحصائية المستخدمة في تلخيص النصوص:

تردد الكلمات: افترضت الطرائق التي تعتمد على تردد الكلمات Word Frequency أن الأجزاء المهمة من النص تحتوي على الأرجح على كلمات تحدث بتردد عالٍ.

الموقع: وظفت أنظمة التلخيص خاصية موقع الجمل Position ضمن المستند. حيث تؤخذ هيكلية المستند عادة في الحسبان عند إعطاء درجات ترتيب للجمل. وفي بعض الطرائق، تفترض أهمية للجمل التي تظهر في بداية المستند أكثر من باقي الجمل. قدم (Baxendale., 1958) تقنيات آلية للتلخيص تعتمد موقع الجمل، واعتمد بأن أفضل موقع للمحتوى المهم في الأجزاء الأولى والأخيرة من الفقرة.

العبارات البارزة : تفترض هذه الخاصية أن وجود بعض العبارات البارزة Cue words and phrases أو غيابها مثل: "كنتيجة لذلك" و "كمثال" و "كحقيقة" تشير إلى أهمية الجمل التي تتضمن هذه العبارات.

التداخل مع العنوان أو الاستعلام: تعين في طريقة التداخل مع العنوان أو الاستعلام Overlap with title or query الكلمات أو العبارات الموجودة في عنوان المستند أو استعلام المستخدم، وتحدد الجمل التي تحوي الكلمات الواردة في العنوان أو الاستعلام. وتأخذ ثقلاً أكبر من باقي الجمل.

2-6-2 تقنيات التعلم الآلي:

الفكرة وراء التعلم الآلي Machine learning هو استخدام مجموعة بيانات تدريب من أجل تدريب نظام التلخيص، حيث ينمذج التلخيص في هذه التقنيات كمسألة تصنيف، تصنف فيها الجمل إلى مجموعتين هي: جمل ملخص وجمل ليست جمل ملخص اعتماداً على خصائص إحصائية بوجود مجموعة تدريب (Oak, 2016) يقدر احتمال اختيار الجملة كجملة ملخص بحسب مستند التدريب والملخص المتولد (Othman et al., 2014). من طرائق التعلم الآلي المستخدمة من أجل تلخيص النصوص:

تقنية نيف بايس Naïve Bayes method: هي طريقة من طرائق التعلم الآلي بمشرف. وتستخدم في التلخيص الآلي بوصف اختيار الجملة فيها مشكلة تصنيف، حيث توضع كل جملة في صنف ثنائي لتحديد فيما إذا كانت ستضمن في الملخص أم لا.

الشبكات العصبونية Neural Network: تستخدم الشبكات العصبونية بوصفها نموذجاً حاسوبياً في علم الحاسوب ومجالات بحث أخرى من أجل حل المشاكل باعتماد طرائق تعلم الآلة. كما استخدمت في مجال التلخيص

الآلي من خلال ثلاثة أطوار: تدريب الشبكة العصبونية، وتطبيق الخصائص، واختيار الجمل. يحدد طور التدريب نوع الجمل التي يجب أن توجد في الملخص. وبعد تدريب الشبكة العصبونية تولد الخصائص المهمة التي يجب أن تضمن في جمل الملخص، وبعد تدريب الخصائص وتوليدها تستخدم الشبكة لاختيار الجمل المهمة لتضمن في الملخص.

أشجار القرار Decision Trees: أشجار القرار هي أدوات لدعم القرارات وهي من طرائق التعلم الآلي بمشرف، لها هيكلية شجرية شبيهة بالمخطط الانسيابي، تمثل كل عقدة داخلية فيها «اختبار» للخاصية (مثال الوجه الظاهر للعملة المعدنية هو طرة أم نقش)، ويمثل كل فرع مخرجات الاختبار، وتمثل كل عقدة نهائية القرار المتخذ بعد احتساب جميع الخصائص. كما يمثل المسار من الجذر إلى العقدة النهائية قواعد التصنيف. اعتمدت بعض نظم تلخيص أشجار القرار في خطوة للابتعاد عن افتراض كون الخصائص المستخدمة في التلخيص مستقلة عن بعضها بعضاً.

نماذج ماركوف المخفية Hidden Markov Models HMM: نموذج ماركوف المخفي هو نوع من أنواع النماذج الإحصائية يُمثل بيانياً في سلسلة ماركوف، تُستخدم نماذج ماركوف المخفية في مهام التعلم الآلي، وقد وُظفت هذه النماذج في أنظمة التلخيص الآلي من أجل الأخذ بالحسبان الاعتمادية بين الجمل ضمن المستند، وذلك بإنشاء نظام قائم على الخاصية التتابعية مع حالات مخفية لها احتمال توزيع على جميع المخرجات الممكنة.

نماذج اللوغاريتم الخطية Log-Linear Models: استخدمت هذه النماذج في التلخيص الآلي، وذلك بتحويل نظام المعادلات، عبر إيجاد لوغاريتم المتغيرات كلها، وتقريب النتائج بمعادلات خطية.

المنطق العائم Fuzzy logic: المنطق العائم هو منطق متعدد القيم، وهو توسيع للمنطق البولياني الذي يوصف القيم البينية بين قيمتين منفصلتين مثل "الواحد" و"الصفر"، و"عال" و"منخفض". والفائدة من المنطق العائم هو توافقيته مع العالم الحقيقي، الذي هو عالم ليس عالم القيمتين. مثلاً عند توصيف حالة الطقس تستخدم صفات عدة مختلفة في المنطق العائم مثل: بارد، بارد جداً، حار، حار جداً، بدلاً من استخدام قيمتين فقط في المنطق البولياني. استخدم المنطق العائم في معالجة اللغات الطبيعية، وكان التركيز على تمكين الحاسوب من فهم اللغة

البشرية التي كانت وراء الواحد والصفر ولم تكن قابلة للتنفيذ بالمنطق البوليني. وقد استخدم مفهوم المنطق العائلي في أنظمة تلخيص النصوص بوصفها فرعاً من فروع معالجة اللغات الطبيعية، للمساعدة في استخراج الجمل التي تتنقل بقيم بين الصفر والواحد اعتماداً على خصائصها، تحدد هذه القيم درجة أهمية تضمين الجمل أو غيابها في الملخص النهائي.

2-6-3 التقنيات الدلالية:

تهتم التقنيات الدلالية Semantic based بمعنى الكلمات والعلاقات بين الكلمات والعبارات والجمل لبناء المفاهيم المقصودة من النص، باستخدام أنطولوجيا أعلامات نوع الكلام والتحليل القواعدي واختيار الجمل ذات المعنى لتوليد الملخص (Chandra et al., 2011). ولا تعتمد الخصائص الإحصائية غير الفعالة التي تعتمد على تنسيق خاص ومكتوبة بنمط المستندات (Youngjoong et al., 2008) مثل كلمات العنوان المشكلة أو الكلمات المكررة. ومن طرائق التقنيات الدلالية مايلي:

التجميع Clustering: من التقنيات الدلالية المستخدمة لتلخيص النصوص آلياً التي تجمع الجمل في مجموعات بالاعتماد على أوجه التشابه (Hatzivassiloglou et al., 2001)، وتختار جملة من كل مجموعة لتوليد الملخص المستخرج.

البيان Graph-based: يستخدم البيان كتقنية دلالية يمثل فيها المستند ببيان غير موجه تمثل الجمل بعقد، وتمثل الحواف بين العقد المتصلة علاقة التشابه بين الجمل. تُعدّ الجملة مهمة إذا كانت مرتبطة ارتباطاً وثيقاً بكثير من الجمل الأخرى (Ramesh et al., 2014).

التحليل الدلالي الكامن Latent Semantic Analysis (LSA): هو تقنية من أجل إيجاد العلاقات المخفية بين المستندات والمقاطع النصية على أساس استخداماتها السياقية.

نظرية الخطاب Discourse theory: استخدمت نظرية الخطاب بوصفها تقنية دلالية تمثل فيها النصوص بهيكل بطريقة ترتبط فيها وحدات الخطاب مع بعضها بعضاً، بدلاً من التعامل مع النص على أنه تسلسل كلمات

وجمل. وتتضمن نظرية الخطاب نظريتين هما: نظرية البنية البلاغية (RST) Rhetorical structure theory التي تستخدم معارف لغوية لتحليل النص، ونظرية البنية الخطابية المجزأة Segmented Discourse Representation Theory (SDRT) تستخدم فيها الرسوم البيانية لنمذجة هياكل الخطاب ويحدد علاقات الترابط من خلال تأثيراتها الدلالية.

الأنطولوجيا Ontology: استخدمت الأنطولوجيا في أنظمة التلخيص الآلي من أجل تحسين جودة التلخيص من خلال استخدام المفاهيم والعلاقات الدلالية بينها.

2-6-4 الذكاء الجماعي:

الذكاء الجماعي Swarm Intelligence هو أداة تعتمد سلوك التجميع لحل عملية معقدة بالعمل الجماعي (Natesh et al., 2016). حفز السلوك الجماعي للنحل والنمل وغيرها الباحثين لاكتشاف نمط حياتهم. الذكاء الجماعي يعتمد على محاكاة الحاسوب لتكرار التفاعل الخلاق لمجموعة مع بعضهم بعضاً ومع بيئتهم لحل مشكلات (Ajith et al., 2008) مثل: الأمثلية Optimization، والروبوتك Robotic، والتوجيه Routing، والتتقيب عن المعلومات Data Mining، والتجميع Clustering وغيرها.

ومن خوارزميات الذكاء الجماعي المعتمدة في تلخيص النصوص:

خوارزمية حشد الجسيمات Particle Swarm Optimization (PSO): هي خوارزمية تطويرية اقترحت من أجل إيجاد الحل الأمثل (Kennedy et al., 1995). تعتمد هذه الخوارزمية على الحركة الجماعية للجسيمات وتبدأ بانتشار الجسيمات لاكتشاف فضاء البحث، وكل فرد في الانتشار له موقع عشوائي وسرعة تضبط ديناميكياً، ليس بالاعتماد فقط على المعلومات الناتجة عن الخبرة الذاتية، ولكن أيضاً بالاعتماد على معرفة جيرانه (Izakian et al., 2015). يتم إيجاد الموقع الأفضل من قبل الجسيمات كلها وتحفظ في الذاكرة، ويُحدَّث الموقع التالي والسرعة لكل جسيم للوصول إلى أفضل قيمة.

خوارزمية بحث الوقواق (Cuckoo Optimization Algorithm (COA): هذه الخوارزمية هي خوارزمية

تطويرية اقترحت من قبل (Yang et al., 2009) من أجل إيجاد الحل الأمثل. وهي مستوحاة من سلوك طير الوقواق في وضع البيض. فمن أجل زيادة فرص بقاء البيضة، تضع طير الوقواق بيوضها في عش طيور أخرى، تكتشف بعض البيوض من قبل الطيور المضيفة ويتخلص منها. والبيض المشابه لبيض المضيف من جانب آخر سوف يكون له فرصة أن يفقس ويكبر، وبعدها تنمو الفراخ، وتصبح طيور الوقواق ناضجة، تقوم طيور الوقواق الناضجة بتكوين بعض المجتمعات، والبحث عن فضاء لإيجاد أفضل بيئة للتكاثر والاستنساخ، وأخيراً يمكن إيجاد الموقع الذي يمكن أن يعيش فيه أكبر عدد ممكن من طيور الوقواق (Lashkari et al., 2017)، ويعرف هذا الموقع على أنه الحد الأعلى العام للوظيفة المطلوبة.

خوارزمية بحث البكتريا (Bacterial Foraging Optimization (BFO): هذه الطريقة مستوحاة من

السلوك الجماعي لبكتريا في الجسم، وهي مقدمة من قبل (Passino, 2002)، تتألف هذه الخوارزمية من ثلاث خطوات هي: الانجذاب الكيميائي Chemotaxis: تحاكي هذه الخطوة حركة البكتريا في حالة السباحة والتراجع، والاستنساخ Reproduction: تدمر في هذه الخطوة البكتريا الضعيفة وتنقسم البكتريا القوية إلى بكتريتين وهذا يساعد في توازن انتشار البكتريا، وإلغاء التشتت Elimination-dispersal: يمكن أن تقتل البكتريا في هذه الخطوة ويمكن أن تتحرك إلى أماكن أخرى.

2-7 تقييم ملخصات النصوص:

عند تصميم نظام تلخيص وتطويره من الضروري وجود أداة أو طريقة لتتبع أداء النظام والتغيرات المنفذة عليه. وقد حدد في (Saggion et al., 2002) ثلاثة أساليب أساسية لتقييم الملخص: الأسلوب البشري Human أواليدوي Manual حيث يقوم البشر بالتقييم، والأسلوب الآلي Automatic حيث تنفذ عملية التقييم كلياً بشكل آلي مثل ROUGE (Lin, 2004)، وأسلوب الدمج Mixed أوشبه الآلي Semi-automatic مثل PYRAMID (Nenkova et al., 2004). استخدم التقييم البشري في بعض الدراسات مثل: ((Schlesinger et al., 2008)، ((El-Haj et al., 2013)، ((El-Haj et al., 2011c).

قسم EL-Haj وزملائه (El-Haj et al., 2011a) قياسات التقييم إلى قياسين هما: الشكل Form والمحتوى Contents، حيث يرتبط القياس الأول بهيكلية أو بنية الملخص مثل: التماسك Coherence أو الصحة القواعدية Grammatical correctness، بينما يقيم القياس الثاني محتوى الملخص.

قسم التقييم في (Jones et al., 1996) إلى فئتين هما: التقييم الجوهرى Intrinsic والتقييم الخارجى Extrinsic.

2-7-1 التقييم الجوهرى:

الهدف الأساسى من التقييم الجوهرى Intrinsic Evaluation هو تقييم أداء محتوى الملخص الآلى بمقارنتها مع ملخص مثالى، له مقياسان، هما: تقييم الجودة Quality Evaluation، وتقييم المحتوى Content Evaluation. يحاول من خلال تقييم الجودة فحص الملخص وخلوه من أخطاء قواعدية أو معلومات متكررة ولها بنية مترابطة، بينما تُقسم مقاييس تقييم المحتوى إلى مجموعتين هما: مقاييس معتمدة على الاختيار المشترك Co-Selection، ومقاييس معتمدة على المحتوى Content-based.

2-7-1-1 المقاييس المعتمدة على الاختيار المشترك

تحدد المقاييس المعتمدة على الاختيار المشترك Co-Selection عدد الجمل المثالية المحتواة ضمن الملخص الآلى. وتتألف من الدقة Precision والاستدكار Recall وقياس F (Steinberger et al., 2009).

- **الدقة:** تقيس الدقة مقدار صحة Correction الملخص أو مقدار صحة المعلومات الموجودة في الملخص، تحسب بتقاطع الجمل المستخرجة والجمل المثالية مقسوماً على الجمل المستخرجة كلها، تعطى وفقاً للعلاقة (1-2):

$$Precision = \frac{\text{relevant Sentences} \cap \text{Retrieved Sentences}}{\text{retrieved Sentences}} \quad (1 - 2)$$

- **الاستدكار:** يستخدم الاستدكار لقياس الاستكمال Completeness أو مقدار كون المعلومات الموجودة في الملخص تغطي المعلومات المهمة في المرجع المرتبط بالملخص، يحسب بتقاطع الجمل المستخرجة والجمل المثالية مقسوماً على الجمل المثالية الموائمة كلها، يعطى وفقاً للعلاقة (2-2).

$$Recall = \frac{Relevant\ Sentences \cap Retrieved\ Sentences}{relavant\ Sentences} \quad (2 - 2)$$

حيث :

- Relevant Sentences : الجمل المثالية الموائمة.
- Retrieved Sentences : الجمل المسترجعة.
- Relevant Sentences $\cap$ Retrieved Sentences : تقاطع الجمل المثالية الموائمة مع الجمل المسترجعة.

ولمقياسي الدقة والاستدكار علاقة عكسية حيث يمكن زيادة الاستدكار مع تقليل الدقة والعكس.

- **المقياس F-Score :** يدمج هذا المقياس بين الدقة P والاستدكار R (Salton, 1989). وهو متوسط مقياسي الدقة والاستدكار، يمكن حساب F-Score وفقاً للعلاقة رقم (2-3).

$$F = \frac{(1 + \beta^2)PR}{\beta^2(P + R)} \quad (3 - 2)$$

حيث β هو متحول مثقل قابل للتغيير لضبط تأثير الدقة والاستدكار.

2-1-7-2 المقاييس المعتمدة على المحتوى:

من أنواع المقاييس المعتمدة على المحتوى Content-based في تقييم التلخيص الآلي:

- **تشابه التجيب Cosine similarity :** يقاس تشابه التجيب من أجل شعاعين X و Y بقياس الزاوية بينهما فيفرض أن X هو شعاع ملخص النص الآلي، و Y هو شعاع الملخص المرجعي، يحسب تشابه التجيب بين الملخصين وفقاً للعلاقة رقم (2-4) .

$$\cos(X, Y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i (x_i)^2} \sqrt{\sum_i (y_i)^2}} \quad (4-2)$$

- **تداخل الوحدات Unit overlap**: يحسب التداخل (Saggion et al., 2002) بين وحدتين نصيتين (كلمات، عبارة، أو أي وحدات نصية) وفقاً للعلاقة (5-2):

$$Overlap(X, Y) = \frac{\|X \cap Y\|}{\|X\| + \|Y\| - \|X \cap Y\|} \quad (5-2)$$

حيث X و Y هما ملخص النص الآلي والملخص المرجعي على الترتيب.

- **مقياس Recall-Oriented Understudy for Gisting Evaluation (ROUGE)**: يقيم جودة الملخص

بمقارنتها مع الملخص البشري K وهناك خمسة مقاييس للتقييم في أداة ROUGE (Lin, 2004) هي:

- **ROUGE-N (N-gram CoOccurence Statistics)**: يعتمد عدد مجاورات n-gram. حيث N يتراوح بين 1 و 4.
- **ROUGE-L (Longest Common Subsequence LCS)**: قياس يعتمد على أطول سلسلة مشتركة.
- **ROUGE-W (Weighted Longest Common Subsequence)**: هو LCS مثقلة.
- **ROUGE-S (Skip-Bigram Co-Occurrence Statistics)**: يستخدم أزواج الكلمات في ترتيب الجمل متجاهلين الفجوات.
- **ROUGE-SU**: يستخدم skip-bigram و Unigram.
- تحسب ROUGE-N وفقاً للعلاقة رقم (6-2).

$$Rouge - n = \frac{\sum_{C \in RSS} \sum_{gram_n \in C} Count_{match}(gram_n)}{\sum_{C \in RSS} \sum_{gram_n \in C} Count(gram_n)} \quad (6-2)$$

حيث:

- **RSS**: الملخص المرجعي.
- **Countmatch(gramn)**: العدد الأعظمي للـ n-gram المتواردة والمشاركة بين الملخص المولد والملخص المرجعي.
- **Count(gramn)**: عدد الـ n-gram في الملخص المرجعي.

2-7-2 التقييم الخارجي:

يعرف قياس التقييم الخارجي Extrinsic Evaluation على أنه مقياس موجه بالتطبيق. ومن أمثلة التطبيقات التي تستخدم في تقييم الملخصات: السؤال-الإجابة، واسترجاع المعلومات. الهدف من هذا التقييم هو تقييم أداء الملخص استناداً إلى تطبيق معين.

الفصل الثالث

الدراسات السابقة

Related Works

3-1 مقدمة:

عرضت في الفصول السابقة أنواع التلخيص الآلي وأهم التقنيات المستخدمة فيها وطرق تقييمها، يقدم هذا الفصل عرضاً لأهم المشاريع وأحدثها ودراسات التلخيص الآلي للنصوص عموماً، ومشاريع التلخيص الآلي للنصوص العربية خصوصاً ودراساتها.

3-2 أهم مشاريع التلخيص الآلي للنصوص ودراساتها:

نشرت أول دراسات بحثية في مجال التلخيص منذ أكثر من ستين عاماً مضت في (Luhn, 1958) حيث استخرجت ملخصات من مستند واحد واستخدام خصائص إحصائية مثل تردد الكلمات، وفي منتصف التسعينات وجد ما يسمى بالتلخيص متعدد المستندات في (McKeown et al., 1995)، والذي أصبح أساسياً مع النمو والتزايد الأسّي للمستندات في الشبكة العنكبوتية. وكمثال على التلخيص متعدد المستندات GoogleNews، وهو أحد الأنظمة المختلفة التي تدمج مجموعات الأخبار المرتبطة مع بعضها والمكتوبة بلغات عدة وتقدمها إلى المستخدمين في ملخص واحد. على الرغم من الفائدة العظيمة فإن هناك تحديات واهتمامات جديدة تجعل عملية التلخيص متعدد المستندات عملية أكثر صعوبة وتعقيداً مقارنة مع التلخيص أحادي المستند. وكأمثلة على هذه التحديات: ترتيب الوحدات النصية للملخص، ومعالجة نسبة الضغط الكبيرة، والحد من التكرار (Lloret et al., 2013). وتغير مجال التلخيص الآلي عبر الزمن بالاعتماد على المتطلبات والتحديات الجديدة. وظهرت أنواع متعددة من التلخيص مثل: التلخيص متعدد المنشورات multi-post (Sharifi et al., 2014)، هو نوع من الملخصات التي تلخص مجموعة المنشورات، مثل: منشورات تويتر Twitter Posts بتحديد المنشور الأكثر أهمية. ووجدت أنواع أخرى من الملخصات

مع وجود متطلبات اجتماعية جديدة، مثل تلخيص التحديثات Update Summaries، حيث تكون لدى المستخدمين معرفة مسبقة بالموضوع، والملخصات الشخصية Personalized Summaries حيث يبني الملخص اعتماداً على الملف الشخصي للمستخدم، وملخصات المشاعر sentiment-based summaries، وملخصات الآراء opinion summaries، من أجل لتوضيح وجهات نظر المستخدمين. كما زاد الاهتمام بالتلخيص متعدد اللغات وخاصة مع أعمال التلخيص متعدد اللغات في دليل TAC 2011 MultiLing Pilot، وورشة العمل ACL 2013 MultiLing. وقد حدد (Lloret et al., 2012) الاتجاهات المستقبلية للتلخيص الآلي. مثل الحاجة إلى تلخيص آلي لتوليد ملخصات متعددة اللغات ومتعددة المستندات، وذلك بسبب النمو الأسّي للمستندات المتاحة بلغات عدة. بالإضافة إلى أن أنواع ملخصات مثل: ملخصات الرأي، والمشاعر، والملخصات الشخصية، ستصبح مهمة بسبب التطور الضخم في شبكات التواصل الاجتماعي. وفي النهاية سوف تصبح الحاجة إلى ملخصات تجريدية إحدى التحديات في هذا المجال.

يمكن العثور على عدد كبير إلى حد ما من أنظمة التلخيص الآلي كأعمال سابقة. تعالج كثيراً من هذه الأنظمة مشكلة التلخيص اعتماداً على نوع الملخصات المطلوبة. وفيما يلي نعرض أهم المشاريع ودراسات التلخيص الآلي للنصوص مصنفة حسب تقنية التلخيص المستخدمة فيها وفقاً لتصنيف (Nazari et al., 2019) الموضح في الشكل رقم (3). كما يسلط الضوء على مزايا كل تقنية وعوائقها. يبين الجدول (1-9) أهم المشاريع والدراسات للتلخيص الآلي للنصوص (في الملحق A).

3-2-1 التقنيات الإحصائية:

تعد الطرائق الإحصائية سهلة التنفيذ، لأنها فقط تستند على الخصائص الفيزيائية للنص، ويمكن استخدامها لتحسين اختيار الجمل المهمة وإلغاء التكرار. ولكنها تقشل في فهم النص، لأنها لا تهتم بمعنى الجمل والكلمات، وهذا يمكن أن يسبب ملخصات بجودة منخفضة. نبين فيما يلي بعض الخصائص الإحصائية وعدداً من الدراسات والأبحاث القائمة على التلخيص الآلي التي استخدمت تلك الخصائص.

تردد الكلمات:

استخدم لوهن (Luhn, 1958) من شركة IBM تردد الكلمات في تطوير أول ملخص آلي اعتمد طريقة تردد الكلمات في تلخيص مقالات علمية. حدد في مشروع (Baxendale, 1958) عتبة مع إسناد ترددات للكلمات أعلى من العتبة. وفي أواخر الثمانينيات اقترح في (Salton, 1989) أن علاقة المصطلح بالمستند تتناسب عكساً مع عدد المستندات الموجودة في المجموعة التي تحتوي هذا المصطلح، وبناء عليه ظهر في هذا المجال قياس يسمى تردد المستند المعكوس (Inverse Document Frequency (IDF، لذا عكست علاقة تردد المصطلح TF لتضمن تردد المستند المعكوس IDF وتغيرت لتصبح $TF \times IDF$ ، وحسبت درجة ترتيب الجملة بمجموع درجات المصطلحات، وبعدها يولد الملخص ببساطة بتجميع الجمل ذات الترتيب الأعلى. كما اعتمد تردد الكلمات في مشاريع عديدة لتلخيص النصوص مثل: (Fachrurrozi et al., 2013).

الموقع:

قدم (Baxendale., 1958) تقنيات آلية للتلخيص تعتمد موقع الجمل، واعتمد أن أفضل موقع للمحتوى المهم في الأجزاء الأولى والأخيرة من الفقرة. استخدم (Edmundson, 1969) طريقة من أجل استخراج الجمل المهمة تعتمد على دمج خاصية الموقع في النص مع خصائص العبارات البارزة وكلمات العنوان. ووجد أن الجمل المواضيعية تميل إلى الظهور مبكراً أو متأخراً في المستندات. استخدمت خاصية موقع الجملة في عديد من أنظمة التلخيص مثل: ((Brandow et al., 1995)، (Teufel et al., 1997)).

العبارات البارزة:

عينت في المشروع (Edmundson, 1969) بعض الكلمات التي تشير إلى مكافأة، على أنها كلمات تجعل الجملة مهمة. افترض في عمل (Pollock et al., 1975) نظاماً سمي ADAM من أجل انتخاب الجمل في الملخصات باعتماد وجود عبارات معينة. عينت في (Teufel et al., 1997) قائمة من العبارات البارزة يدوياً لتحديد أهمية الجملة. بعدها أتمت في (Aone et al., 1998) عملية إنشاء قوائم الكلمات والعبارات البارزة في خصائص

مدربة متعلمة من ملخصات أمثلة. واستخدمت أنظمة أخرى هذه الخاصية مثل ((Cristea et al., 2003)، (Jaoua (Wan et al., 2007، et al., 2003)).

التداخل مع العنوان أو الاستعلام:

من الأنظمة التي اعتمدت هذه الخاصية مع غيرها من الخاصيات ((Conroy et al., 2001)، (Hovy (et al., 1998)).

3-2-2 تقنيات التعلم الآلي:

تعد تقنيات التعلم الآلي فعالة في تعلم الخصائص الأساسية لعمل الملخص، وينتج عنها نتائج تلخيص جيدة إلا أنها تعتمد على تعقيد قواعد بيانات التدريب في الحساب، والتي تختلف من لغة إلى أخرى، وليست ثابتة من أجل المستندات كلها. ومن طرائق التعلم الآلي المستخدمة من أجل تلخيص النصوص:

نيف بايس :

ومن مشاريع التلخيص الآلي التي استخدمت هذه التقنية: ((Kupiec et al., 1995)، (Daume et al., 2005)، (Dorr et al., 2003)، ((Sarkar et al., 2012).

الشبكات العصبونية:

ومن المشاريع التي اعتمدت الشبكات العصبونية في تلخيصها: ((Kaikhah, 2004)، (Svore et al., 2007)، (Burges et al., 2005)، ((Sarda, 2015).

أشجار القرار:

استخدمت أشجار القرار في مشاريع تلخيص عديدة منها: ((Lin, 1999)، (Chen et al., 2009)، (Jin (et al., 2008)).

نماذج ماركوف المخفية:

وظفت أنظمة متعددة نماذج ماركوف المخفية في أنظمة التلخيص الآلي منها (Schwartz et al., 1997)، (Luhn, 1958)، (Zajic et al., 2002)، (Conroy et al., 2005).

نماذج اللوغاريتم الخطية:

طبقت نماذج اللوغاريتم الخطية في مشاريع عديدة مثل: ((Youngjoong2 2008)، (Hatzivassiloglou et al., 2001)، (Titov et al., 2008)).

المنطق العائلي:

ومن أنظمة التلخيص الآلي للنصوص التي طبقت هذه التقنية: ((Kyoomarsi et al., 2008)، (Hannah et al., 2011)، (Patil et al., 2014)، (Thorat, 2014)).

3-2-3 التقنيات الدلالية:

ساهم استخدام هذه التقنيات في التلخيص الآلي للنصوص في تحسين الجودة على نطاق واسع من خلال إنشاء ملخصات أكثر تماسكاً وأقل تكراراً وأكثر إفادة. ومع ذلك، فهي مهمة صعبة نظراً لأنها تواجه صعوبات في استخدام أدوات التحليل الدلالي عالية الجودة والموارد اللغوية، لأنها تتطلب ذاكرة لحفظ المعلومات الدلالية، مثل: الأنطولوجيا، وسعة المعالج بسبب المعالجة اللغوية المعقدة. وهناك طرائق متعددة طُوّرت لتوليد ملخصات باعتماد التقنيات الدلالية هي:

التجميع:

على الرغم من استخدام تقنيات التجميع لتقليل تكرار البيانات، إلا أن الملخص المنشأ باستخدامها قد يفتقر إلى المعنى بما يكفي، لأن الجمل المختارة تُصنف بشكل أساسي اعتماداً على القرب من النقطة الوسطى العنقودية دون الأخذ بالمعاني (Nenkova et al., 2011). هناك مشاريع تلخيص آلي عديدة اعتمدت أيضاً التجميع مثل: ((Sarkar, 2009)، (Anjali et al., 2013)، (Bazghandi et al., 2012)، (Pei-ying et al., 2009)).

البيان:

إن استخدام التقنيات المعتمدة على البيان له مساهمة إيجابية في التلخيص متعدد المستندات لقدرته على اختيار موضوعات متميزة من البيانات الفرعية غير المتصلة. ومع ذلك، فإن إنشاء البيانات الفرعية اعتماداً على مقاييس التشابه الإحصائي دون إيلاء أي اهتمام لمعنى النص، يخاطر بإنتاج ملخص أقل إفادة (Lagrin et al., 2017). وقد طبق نظام للتلخيص باعتماد البيان في كثير من المشاريع منها: ((Natesh et al., 2016)، (Moawad et al., 2012)، (Kavita et al., 2010)، ((Niraj et al., 2016).

التحليل الدلالي الكامن:

استخدم التحليل الدلالي الكامن في نظم تلخيص متعددة، كتدبير للتغلب على مشكلة الكلمات المترادفة في تمثيلات المستندات والاستعلامات. ومن هذه الأنظمة: ((Yeh et al., 2005)، (Steinberger et al., 2004)، ((Brown, 2010).

نظرية الخطاب:

استخدمت نظرية الخطاب كتقنية دلالية، وتمثل فيها النصوص بهيكل بطريقة ترتبط فيها وحدات الخطاب مع بعضها بعضاً، وتتضمن نظرية الخطاب نظريتين هما: نظرية البنية البلاغية التي استخدمت هذه النظرية لتلخيص النصوص في مشاريع عديدة مثل: (Chengcheng, 2010)، ونظرية البنية الخطابية المجزأة التي استخدمت في التلخيص الآلي للنصوص في مشاريع عديدة منها: (Zhengyuan et al., 2019).

الأنطولوجيا:

استخدمت الأنطولوجيا في أنظمة التلخيص الآلي من أجل تحسين جودة التلخيص واستخدام علاقاتها الدلالية، ومن الأنظمة التي استخدمت فيها هذه التقنية: ((Brazilay et al., 2015)، (Saxena et al., 2016)) حيث ولدت السلاسل المعجمية بالاستعانة بأنطولوجيا شبكة الكلمات WordNet لتحديد الكلمات المرتبطة دلاليًا، واختيار السلسلة المعجمية الأقوى، كما كانت الأنطولوجيا أساس اعتمد عليه في (Ragunath et al., 2015) لتحسين جودة التلخيص.

3-2-4 اعتماد الذكاء الجماعي:

استخدمت أنظمة التلخيص خوارزميات معتمدة على الذكاء الجماعي، فقد استخدمت خوارزمية تحسين الجسيمات في تلخيص النصوص من قبل (Binwahlan et al., 2009) من أجل إعطاء درجات ترتيب للجمل لتوليد الملخص مع مطابقة عليا مع الملخصات البشرية. واستخدمت خوارزمية بحث الوقواق في تلخيص النصوص من قبل (Mirshojaei et al., 2015). كما اعتمدت خوارزمية بحث البكتريا في أتمتة تلخيص النصوص من قبل (Nikoo et al., 2012). تتميز تقنيات الذكاء الجماعي بأداء جيد إلا أنها تتطلب تعقيداً في القياس.

3-3 الأبحاث والتقنيات المستخدمة لتلخيص النصوص العربية

يوجد كثير من الأبحاث والدراسات والأنظمة الخاصة بالتلخيص الآلي للنصوص باللغات الأوروبية المتعددة وغيرها من اللغات العالمية والتي شرحت في الفقرة السابقة، إلا أن الأبحاث والدراسات في مجال تلخيص النصوص العربية ضئيلة وتتمو ببطء. سوف نسلط الضوء فيما يلي على أهم الأبحاث والدراسات الحديثة في مجال تلخيص النصوص العربية مصنفة حسب تقنية التلخيص المستخدمة فيها. يبين الجدول (9-2) (في الملحق A) أهم دراسات تلخيص النصوص العربية ومشاريعها.

3-3-1 التقنيات الإحصائية:

يوجد عديد من مشاريع تلخيص النصوص العربية ودراساتها التي تعتمد على التقنيات الإحصائية التي استخدمت مجموعة من الخصائص لتقييم أهمية الجمل إحصائياً منها: ((El-Ghannam et al., 2013)، (El- (Haj et al., 2013)، (Gupta et al., 2011)، (Ferreira et al., 2012)، (Abdelkrime et al., 2015)، (Litvak (et al., 2016)، (El-Shishtawy et al., 2012)، (Alotaiby et al., 2012)).

3-3-2 تقنيات التعلم الآلي:

وُظفت بعض أنظمة تلخيص النصوص العربية التعلم الآلي Machine Learning باعتماد خوارزميات عدة للتعلم الآلي مثلاً: كامل الباحثون في (Sobh, 2009) تقنية بايزن مع البرمجة الجينية، ودمج الباحثون تقنيات عدة

للتعلم الآلي في (Fattah et al., 2009)، وهي الخوارزمية الجينية، والتراجع الرياضي، والشبكات العصبونية، ونموذج غوسن، من أجل عملية التلخيص الآلي للنصوص العربية لمقارنتها، واعتمد أيضاً ملخص (El-Fishawy et al., 2013) التعلم الآلي باستخدام النموذج التراجعي وأشجار القرار من أجل تلخيص منشورات متعددة Multi-post. كما قدم نظام (Belguith et al., 2014) نظام تلخيص دمج فيه بين تقنيتي التحليل البياني والتعلم الآلي بمشرف. استخدمت الخوارزميات الجينية في ((Yazan et al., 2016)، (Abuobieda et al., 2012)، (Tanfoury et al., 2021)) من أجل تلخيص النصوص. واعتمد في ((Boudabous et al., 2010)، (Belkebir et al., 2015)) على تقنيات التعلم الآلي بمشرف، كما اقترح الباحثون في (Qaroush et al., 2021) طريقة تلخيص آلية بالدمج بين الخصائص الإحصائية والدلالية واعتماد تقنية التعلم الآلي لإنتاج الملخص. استخدمت تقنيات التعلم العميق Deep learning في مشروع (Al Maleh et al., 2020) باعتماد الشبكات العصبونية لتلخيص النصوص العربية.

3-3-3 التقنيات الدلالية:

استخدمت تقنيات دلالية متعددة في تلخيص النصوص العربية منها: التحليل الدلالي الكامن، واستخدام الأنطولوجيا، والتجميع، والبيان، ونظرية الخطاب. فقد استخدم التحليل الدلالي الكامن في: ((Brazilay et al., 2015)، ((Froud et al., 2013)). كما استخدمت الأنطولوجيا كتلخيص دلالي يأخذ بالحسبان المفاهيم والعلاقات الدلالية بين المصطلحات، كما في نظام OSSAD (Imam et al., 2013)، وهو نظام مقاد بالاستعلام، ويستخدم أنطولوجيا شبكة الكلمات العربية لتوسيع الاستعلام، واستخدمت السلاسل المعجمية التي لها علاقات دلالية في (AL-Khawaldeh et al., 2015). استخدمت تقنية التجميع لتلخيص النصوص العربية في: ((Fejer et al., 2014)، ((Haboush et al., 2012)، ((El-Haj et al., 2011a)). اعتمدت تقنية البيان في أنظمة مثل: ((Erkan et al., 2004)، ((Thomas et al., 2015)). طبقت نظرية الخطاب في كثير من أنظمة تلخيص النصوص العربية مثل: ((Elghazaly et al., 2012)، ((Keskes, 2015)، ((Ibrahim et al., 2013)).

3-3-4 تقنيات الذكاء الجماعي:

استخدمت تقنية الذكاء الجماعي Swarm intelligence باعتماد خوارزمية حشد الجسيمات PSO في

تلخيص النصوص العربية في: (AlZahrani et al., 2015)، (Al-Abdallah et al., 2017).

3-4 أهم أدوات وأنظمة التلخيص الآلي:

هناك عديد من الأنظمة والأدوات المعدة لتلخيص النصوص آلياً، وهي عبارة عن مواقع على شبكة الإنترنت منها ما هو مجاني ومنها ما يتطلب اشتراك، تعتمد غالبيتها على تقنيات إحصائية أو تقنيات التعلم الآلي في تلخيصها للنصوص، ولا تعتمد تقنيات دلالية، وتستخدم غالباً لتلخيص النصوص في اللغات الأوروبية، وبعضها يدعم اللغة العربية. فيما يلي نقدم عرضاً لأهم هذه الأدوات والأنظمة:

: Resoomer

أداة سهلة الاستخدام تساعد على تلخيص النصوص وتحليلها ، يُمكن موقعها <https://resoomer.com/> من تلخيص نص ما عن طريق نسخه ولصقه، مع إمكانية تحديد نسبة التلخيص. وتعمل هذه الأداة بلغات عديدة بما لا تتضمن العربية. هناك نسخة مجانية من هذه الأداة على الإنترنت تحتوي إعلانات ولا تحتاج إلى تسجيل دخول، كما يمكنها تصدير الملخص إلى مستندات بالتنسيقات الشائعة.

: QuillBot

أداة تلخيص أخرى على شبكة الإنترنت <https://quillbot.com/> تتيح أخذ أي محتوى نصي استخراج الأجزاء المهمة منه كملخص. تستخدم هذه الأداة خوارزميات الذكاء الاصطناعي وتعتمد خاصية العبارات المفتاحية.

: TextCompactor

أداة تلخيص مجانية تسمح بتلخيص أي جزء من النص بعد تحديد نسبة التلخيص المطلوبة، تعتمد خاصية توارد المصطلحات والعبارات وتكرارها في تحديد أهمية الجمل. لا يتطلب موقعها الخالي من الإعلانات

<https://www.textcompactor.com> التسجيل لاستخدامه، إلا أنه لا يسمح باستيراد عناوين URL أو الملفات وحفظ الملخص بتنسيقات المستندات الشائعة.

: SummarizeBot

أداة تلخيص موفرة للوقت تقدم ملخصات نصية دقيقة، تستخدم تقنية الذكاء الاصطناعي لتحليل النصوص واستخراج الأفكار الرئيسية للملخص. موقعها <https://www.summarizebot.com> خالٍ من الإعلانات وسهل الاستخدام. ليس هناك قيد على عدد كلمات النص المطلوب تلخيصه ويسمح باستيراد أو تصدير الملفات أو عناوين URL، ويدعم الكثير من اللغات ومن ضمنها اللغة العربية.

: SMMRY

أداة تلخيص تم إنشاؤها لتقديم طريقة فعالة لفهم المقالات والنصوص وتلخيصها. تقوم الأداة بإزالة عبارات الانتقال والجمل غير الضرورية والأمثلة المفرطة. يوفر موقعها <https://smmry.com> تصميماً سهل الاستخدام وإعدادات متقدمة. ويمكن من خلاله تلخيص نصوص إما بنسخ النص ولصقه مباشرة أو لصق عنوان URL .

:Summarizing tool

تعتمد هذه الأداة أيضاً على خوارزمية الذكاء الاصطناعي لفهم المحتوى وإعطاء درجات أهمية للجمل وتوليد الملخص النهائي. كما يتضمن موقعها <https://summarizing-tool.com> مدققاً لغوياً لتدقيق الملخص .

:Tools4noobs

أداة مجانية على الويب توفر من خلال موقعها <https://www.tools4noobs.com/summarize> إمكانية نسخ النص ولصقه مباشرة أو لصق عنوان URL، وتحديد نوع الملخص ونسبة التلخيص المطلوبة أو تحديد عدد الأسطر المطلوبة. كما يحتوي موقعها إعلانات.

:TextSummarization

أداة تلخيص مجانية تستخدم الذكاء الاصطناعي في تلخيصها، وتتيح من خلال موقعها <https://textsummarization.net/text-summarizer> تحديد عدد الجمل المطلوبة في الملخص ، وتوفر إمكانية لصق عنوان URL أو النص.

:AutoSummarizer

تُستخدم هذه الأداة لتلخيص المقالات النصية حيث يمكن استخراج الجمل المهمة وترتيبها بناءً على أهميتها. إلا أنها تقتصر إلى توفير إمكانية استيراد الملفات وتصديرها، كما أن الحد الأعظمي لطول النصوص التي يمكن تلخيصها هو 3375 كلمة على الأكثر. ويحتوي موقعها <https://autosummarizer.com> على إعلانات.

: Free Summarizer

هو من أسرع المواقع <http://freesummarizer.com> التي تقوم بتلخيص النصوص على الإنترنت، الذي يلخص نصوصاً بلغات مختلفة عن طريق لصق النص يدوياً، وتحديد عدد الجمل المطلوبة في الملخص. كما يتضمن أداة تدقيق لغوي للتحقق من صحة الملخص.

:Summarizer

أداة تلخيص مضمنة في متصفح كروم Google Chrome تساعد في تلخيص أي محتوى على الإنترنت. تعالج مقالات وأجزاء مختلفة من نص في متصفح الويب، مع تضمين سبعة بالمئة فقط من المقالة الأصلية في الملخص الناتج. لكنها تعمل فقط مع متصفح الكروم وتلخيص نصوص باللغة الإنكليزية فقط، ولا يمكن تحميل الملخص الناتج أو استيراد / تصدير ملف لتلخيصه.

: Split Brain Summary

تساعد هذه الأداة في تلخيص نصوص بتسع وثلاثين لغة مختلفة لانتضمن اللغة العربية، كما يمكن من خلال موقعها <https://www.splitbrain.org/services/ots> إدخال عنوان URL بدلاً من نسخ النص ولصقه

في الأداة. وتحديد نسبة التلخيص المطلوبة، تقتقر هذه الأداة إلى خيار استيراد الملفات أو تصدير الملخص إلى التنسيقات الشائعة، لا يحتوي موقعها على إعلانات.

:AppZaza Article Summarizer

أداة مجانية ولا تتطلب التسجيل. يعمل موقعها <https://appzaza.com/article-summarizer> بشكل أفضل عند تلخيص المستندات جيدة التنظيم فقط. يدخل النص إليه عن طريق القص واللصق أو كتابته فيه، ولا يسمح باستيراد الملفات وتصدير الملخص إلى التنسيقات الشائعة، وهي أداة تدعم اللغة الإنكليزية فقط.

:Paraphrase

أداة مجانية عنوان موقعها على الإنترنت <https://paraphrasingtool.ai> تستخدم تقنية الذكاء الاصطناعي لتلخيص نصوص بلغات عديدة من ضمنها اللغة العربية، عن طريق لصق النص المطلوب تلخيصه في موقعها بما لا يتجاوز 3000 حرفاً بعد تحديد لغة النص المطلوبة.

:Lakhasly

يعد موقع التلخيص <https://lakhasly.com> من أبرز المواقع التي تقوم بتلخيص النصوص الإنكليزية والعربية مجاناً على الإنترنت، تستخدم فيه تقنيات إحصائية لتحديد أهميه جمل النص. ويمكن من خلاله لصق النصوص المراد تلخيصها، وتحديد نسبة التلخيص المطلوبة، وتصدير الملخص الناتج إلى مستندات بالتصنيفات الشائعة.

:Summazer

تطبيق مجاني عنوانه www.summazer.com لتلخيص نص أو صفحة ويب من خلال تمييز الجمل المهمة بلون شفاف. وذلك إما بلصق النص أو كتابة عنوان URL وتحديد اللغة مثل العربية، ونسبة التلخيص المطلوبة. كما يقوم التطبيق بإجراء تدقيق لغوي.

:Almaqalah

يعد موقع المقالة لتلخيص النصوص العربية <https://almaqalah.com/ara-summarizer> من المواقع التي تلخص النصوص العربية، ويتميز بالبساطة وسهولة الاستخدام، يمكن من خلاله كتابة النص المطلوب تلخيصه أولصقه، كما يُمكن من إدخال كلمات مفتاحية للنص المطلوب تلخيصه.

3-5 وصف المشكلة وحلها

اللغة العربية هي اللغة الأم لأكثر من 300 مليون شخص حول العالم، وهي واحدة من اللغات الست الرسمية لدى الأمم المتحدة (Farghaly et al., 2009). يستدعي التعامل معها تحديات كثيرة في مجال معالجة اللغات الطبيعية (Farghaly et al., 2009)، ويتطلب أعمال معالجة مختلفة عند تنفيذ عمليات مثل: الترجمة الآلية والتلخيص الآلي للنصوص العربية.

كما بينت الدراسات قلة الأبحاث والمشاريع والأدوات التي اهتمت بتلخيص النصوص العربية وحداتها مقارنة مع اللغات الأخرى، واتبعت في معظمها الأسلوب الاستخراجي واستخدام التقنيات الإحصائية، وتقنيات التعلم الآلي، وقلما اتبعت الأسلوب التجريدي، واستخدام التقنيات الدلالية، والاعتماد على مصادر معرفية كالأنطولوجيا. كما تبين أن بعض المشاريع مثل: (Imam et al., 2013) اعتمدت على الأنطولوجيا في تلخيصها للنصوص العربية، إلا أنها استخدمت أنطولوجيات مترجمة تقتصر إلى وجود مفاهيم عربية الأصل، واستفيد من الأنطولوجيا في تلخيص آلي مقاد بالاستعلام، وذلك لتوسيع الاستعلام باستخدام العلاقات الدلالية في الأنطولوجيا، ولم يستند من الأنطولوجيا في تلخيص عام غير مقاد بالاستعلام للاستفادة منها في تمثيل النصوص بمفاهيمها وعلاقاتها الدلالية، والاعتماد على هذا التمثيل في عملية التلخيص من خلال تثقيل الجمل وتحديد أهميتها اعتماداً على خصائص مفاهيمها وعلاقاتها الدلالية، مما قد يحسن من جودة التلخيص.

لذا كان لابد من الاستفادة من دراسات التلخيص للنصوص باللغات العالمية الأخرى عموماً وباللغة العربية خصوصاً. وإيجاد طريقة تلخيص آلي للنصوص العربية تراعى فيها خصوصية اللغة العربية وتستخدم تقنيات

دلالية باستخدام مصادر معرفية كالأنطولوجيا العربية في تلخيص عام يمثل النصوص المراد تلخيصها بمفاهيمها وعلاقاتها الدلالية، من أجل تحسين تقييم أهمية الجمل على أساس مفاهيم مصطلحاتها والعلاقات بينها، مما قد يحسن من جودة التلخيص.

ومن الضروري إيجاد أنطولوجيا عربية للاعتماد عليها في عملية تلخيص النصوص العربية، وذلك من خلال توسيع أنطولوجيا شبكة الكلمات العربية Arabic WordNet المترجمة بإغنائها بمفاهيم ومصطلحات وعلاقات مستخرجة من الموسوعة العربية باستخدام تقنية التقابل. مع الأخذ بالحسبان أن شبكة الكلمات العربية هي أكثر الأنطولوجيات العربية استخداماً، والموسوعة العربية هي أغنى مصدر معرفي عربي شامل للمعارف الإنسانية كافة.

الباب الثاني

توليد ملخص آلي للنصوص العربية

الفصل الرابع

نموذج عام للملخص الآلي

General Model for Automatic Summarizer

1-4 مقدمة

عرض في الفصول السابقة الطرائق والتقنيات المستخدمة في التلخيص الآلي للنصوص. كما عرضت الأبحاث والدراسات التي تناولت تلخيص النصوص العربية بوجه خاص. نقدم في هذا الفصل نموذجاً مقترحاً لتلخيص النصوص العربية اعتماداً على الطريقة الدلالية باستخدام الأنطولوجيا.

2-4 نموذج التلخيص الآلي للنصوص العربية

وضع نموذج لتلخيص النصوص وذلك بالاستفادة من دراسات التلخيص للنصوص باللغات العالمية الأخرى عموماً وباللغة العربية خصوصاً، مراعيّاً خصوصية اللغة العربية، ومستخدمّاً تقنيات دلالية باعتماد مصادر معرفية كالأنطولوجيا العربية، في تلخيص عام يمثل النصوص المراد تلخيصها بمفاهيمها وعلاقاتها الدلالية. وقد تضمن هذا النموذج طريقة لتوليد أنطولوجيا عربية للاعتماد عليها في عملية تلخيص النصوص العربية باستخدام تقنية التقابل في توسيع شبكة الكلمات العربية المترجمة، وإغنائها بمفاهيم ومصطلحات الموسوعة العربية وعلاقاتها الدلالية.

يتألف نموذج التلخيص الآلي للنصوص العربية المقترح، كما هو موضح في الشكل (4-1)، من مراحل

عدة وهي:

1- توليد أنطولوجيا عربية وذلك بتوسيع شبكة الكلمات العربية Arabic Wordnet بالتقابل Mapping مع

الخصائص المستخرجة من نصوص الموسوعة العربية.

2- بناء نظام لتلخيص النصوص العربية بطريقة دلالية اعتماداً على الأنطولوجيا العربية المولدة .

3- تقييم نظام التلخيص المقترح.

4- تطبيق هذا النظام على نصوص الموسوعة العربية لتلخيصها.



الشكل (1-4) مراحل نموذج التلخيص الآلي للنصوص العربية

وسيقدم شرح توضيحي عن تلك المراحل في الفقرات التالية من هذا الفصل.

4-2-1 توليد أنطولوجيا عربية:

لا يوجد أنطولوجيا عربية تضم المفاهيم العربية ومجموعة مترادف المصطلحات العربية وعلاقاتها الدلالية إلا في بعض المشاريع البحثية غير المكتملة. وتعد شبكة الكلمات العربية أكثر الأنطولوجيات العربية المتاحة والمتوفرة بوصفها مصدراً معرفياً مفتوحاً يستخدم على نحو واسع، إلا أنها أنشئت مترجمة لشبكة الكلمات الإنكليزية وللمفاهيم الإنكليزية.

لذا حاولنا في هذا البحث إيجاد طريقة لتوسيع شبكة الكلمات العربية آلياً باستخدام تقنية التقابل Mapping مع خصائص الموسوعة العربية التي تعد مصدراً معرفياً غنياً بالمعلومات والمفاهيم العربية في المجالات كافة. ولا سيما أن محاولة بناء أنطولوجيا عربية من العدم يدوياً، تعد مهمة مكلفة من حيث الوقت والجهد والموارد الأخرى.

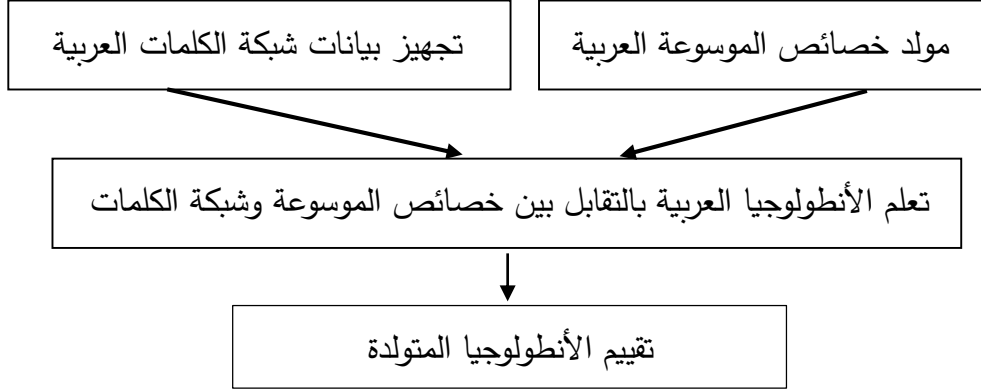
تولد أنطولوجيا عربية بالتقابل Mapping بين نصوص الموسوعة العربية مع شبكة الكلمات العربية Arabic Wordnet. وذلك وفقاً للمراحل الموضحة بالشكل (4-2)، والتي تتضمن:

1- مولد خصائص الموسوعة العربية.

2- تجهيز بيانات شبكة الكلمات العربية.

3- تعلم الأنطولوجيا العربية بالتقابل بين خصائص الموسوعة العربية وبيانات شبكة الكلمات العربية.

4- تقييم الأنطولوجيا المتولدة.



الشكل (4-2): مراحل توليد الأنطولوجيا العربية

يتضمن الفصل السادس شرحاً عن شبكة الكلمات وعلاقاتها، وتعريفاً بالموسوعة العربية، وكيفية توليد

أنطولوجيا عربية بتوسيع شبكة الكلمات العربية بالتقابل مع مفاهيم الموسوعة العربية.

2-2-4 بناء نظام تلخيص الآلي للنصوص العربية:

اقترح نظامان لتلخيص النصوص العربية وهما نظام تلخيص باعتماد الأنطولوجيا ونظام تلخيص إحصائي

لايعتمد الأنطولوجيا، واستخدم الأخير من أجل المقارنة وتقييم النظام المعتمد على الأنطولوجيا. وقد اعتمدت

الأنطولوجيا العربية المولدة بالتقابل بين شبكة الكلمات العربية وخصائص الموسوعة العربية.

يطبق النظامان المقترحيان تلخيصاً استخراجياً عاماً أحادي المستند وأحادي اللغة، ويتشاركان في مراحل

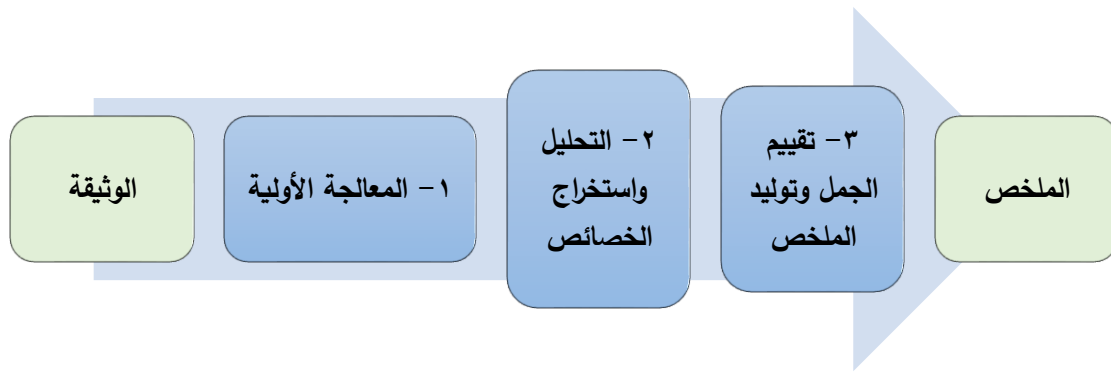
التلخيص الأساسية، ويختلفان في تنفيذ تلك المراحل حيث تستخدم الأنطولوجيا العربية في التلخيص باعتماد

الأنطولوجيا لإعادة المصطلح إلى مفهومه والتعامل مع المفاهيم بدلاً من المصطلحات في تقليل الخصائص

والجمل.

يوضح الشكل رقم (4-3) مراحل التلخيص الآلي في نظامي التلخيص المقترحين وهي:

- 1 المعالجة الأولية Text Preprocessing لمستندات المصدر .
 - 2 التحليل واستخراج الخصائص وصياغتها .Analysis and Feature extraction and formulation.
 - 3 تقييم الجمل وتوليد الملخص Sentence evaluation and summary generation.
- يستقبل نظام التلخيص المستند المطلوب تلخيصه كدخل، ومتغيرات اختيارية يحددها المستخدم وهي نسبة التلخيص ونوع التلخيص، وينتج عن النظام ملخص مستند الدخل كخرج.



الشكل (3-4) مراحل التلخيص الآلي

يُعد المستند في مرحلة المعالجة الأولية ويُقدم بطريقة بنيوية موحدة لتسهيل العمل في المراحل القادمة. تحسب في المرحلة الثانية مجموعة من الخصائص الإحصائية والدلالية لكل جملة لتبين أهميتها، وتستخدم هذه الخصائص في مرحلة تقييم الجمل وتوليد الملخص.

يتضمن الفصل السابع شرحاً عن هذه المراحل وكيفية تنفيذها.

3-2-4 تقييم أنظمة التلخيص المقترحة:

تُستخدم مجموعة الملخصات العربية من (Essex (EASC (El-Haj et al., 2010) لاختبار نظم التلخيص المقترح وتقييمه. ويُطبق التقييم باعتماد التوارد المشترك واستخدام مقاييس الدقة precision والاستدكار recall ومعامل F، وهي المقاييس الأكثر شهرة واستخداماً (Oufaida et al., 2014)، تقيس الدقة مقدار صحة correction الملخص أو مقدار صحة المعلومات الموجودة في الملخص، وتحسب بتقاطع الجمل المستخرجة والجمل الموائمة مقسوماً على الجمل المستخرجة كلها، ويستخدم الاستدكار لقياس الاستكمال completeness

أومقدار كون المعلومات الموجودة في الملخص تغطي المعلومات المهمة في المرجع المرتبط بالملخص. ويحسب بتقاطع الجمل المستخرجة والجمل المثالية مقسوماً على الجمل المثالية الموائمة كلها.

4-2-4 تطبيق نظام التلخيص على نصوص الموسوعة العربية:

يطبق التلخيص الآلي على نصوص من الموسوعة العربية لتلخيصها وتقييم نتائج التلخيص. وذلك

باختيار مقالات من اختصاصات مختلفة وبأطوال مختلفة، لكل بحث ملخص مرجعي لخص من قبل باحثين

متخصصين في هيئة الموسوعة العربية لغرض المقارنة والتقييم.

الفصل الخامس

توليد أنطولوجيا عربية

Generation of Arabic Ontology

1-5 مقدمة

هناك اهتمام كبير في هذه السنوات في الأنطولوجيا Ontology كأداة من أجل تمثيل المعارف. يعد توليد الأنطولوجيا اليوم من الأبحاث الأساسية ذات أهمية في تنفيذ الكثير من التطبيقات مثل البحث الدلالي (Benassi et al., 2004) والتلخيص الآلي (Carenini et al., 2008) وغيرها. ولقد طُورت العديد من المشاريع البحثية والنماذج المخبرية والأدوات من أجل هذا الغرض، إلا أن توليد أنطولوجيا عربية مازال يتطلب جهوداً بحثية كبيرة. ومن أكثر الأنطولوجيات العربية استخداماً شبكة الكلمات العربية (Arabic WordNet (AWN والتي وُلدت بترجمة شبكة الكلمات الإنكليزية إلى العربية، والمنتشرة على نحو واسع كأنطولوجيا معجمية تُستخدم في الأبحاث الحالية بسبب إتاحتها مجاناً وتغطيتها الواسعة. ولكن على الرغم من انتشار استخدامها، إلا أنها تقتصر إلى بعض المعلومات العربية المفيدة لمعالجة اللغة. لذا تم في هذا البحث تطوير طريقة لتوسيع شبكة الكلمات العربية للحصول على أنطولوجيا عربية مناسبة لاستخدامها في التلخيص الآلي للنصوص العربية. وذلك بطريقة التقابل الآلية لمفاهيمها ومصطلحاتها مع خصائص Features مقالات الموسوعة العربية التي تعد مصدراً معرفياً غنياً في مجالات المعرفة الإنسانية كافة. وبناءً عليه أوجدت أنطولوجيا عربية جمعت بين شمولية شبكة الكلمات العربية المترجمة وبين جودة مفاهيم الموسوعة العربية ومصطلحاتها وعلاقاتها العربية. بينت النتائج أنه أُضيف عدد جيد من المفاهيم والمصطلحات العربية والعلاقات الدلالية إلى أنطولوجيا شبكة الكلمات العربية، وعليه يمكن اعتبار الأنطولوجيا المولدة أنطولوجيا عربية جوهرية Core Arabic ontology يعاد استخدامها في التطبيقات العربية الدلالية، ويمكن استكمالها للحصول على أنطولوجيا عربية فعالة.

يُقدم فيما يلي تعريفاً للمصطلحات المستخدمة وهي: الأنطولوجيا، تعلم الأنطولوجيا، شبكة الكلمات العربية، الموسوعة العربية. كما يُعرض شرحٌ مختصرٌ عن أهم الدراسات والأبحاث المستخدمة في توليد الأنطولوجيا العربية، وأخيراً يُسلط الضوء على طريقة التقابل في توليد الأنطولوجيا المستخدمة في هذا البحث.

5-2 تعريف الأنطولوجيا:

تعد الأنطولوجيا أحد أهم مكونات الوب الدلالي Semantic web وحجر الأساس لعمله، وتُعرّف على أنها طريقة لتمثيل المفاهيم من حولنا عن طريق الربط بينها بعلاقات دلالية، مما يساعد على فهم أوسع للمفاهيم المختلفة. بهذه الطريقة يمكن جعل الحاسوب الآلي يصل لمستوى إدراك المعاني قريب من فهم الإنسان وإدراكه. الأنطولوجيا هي "توصيف صوري صريح formal explicit specification للمفاهيمية التشاركية shared conceptualization" (Gómez-Pérez et al., 2004). تشير المفاهيمية Conceptualization إلى توليد النماذج المجردة لبعض الظواهر في العالم بتحديد المفاهيم المرتبطة بتلك الظاهرة. ويُقصد بكلمة الصريح Explicit هو تعريف المفاهيم المستخدمة ومحدداتها على نحوٍ صريح. وتشير كلمة صوري Formal إلى تمثيل الأنطولوجيا بطريقة صورية يمكن قراءتها ومعالجتها حاسوبياً. وترفض التشاركية Shared أن تكون الأنطولوجيا خاصة أوضمنية ببعض الأفراد وإنما معرفة عامة وتشاركية للمجموعة.

تتألف الأنطولوجيا من جانب من المفاهيم concepts، والعلاقات التصنيفية taxonomic relationships التي تحدد العلاقات الهرمية بين المفاهيم، والعلاقات غير التصنيفية non-taxonomic relationships بينهم، ومن جانب آخر حالات المفاهيم concept instances ومحدداتها assertions. يمكن تمثيل الأنطولوجيا بلغات صورية مثل XML و RDF و OWL.

3-5 تعلّم الأنطولوجيا

تعلّم الأنطولوجيا هو عملية التوليد الآلي أو الشبه الآلي للأنطولوجيا، متضمناً استخراج المصطلحات المرتبطة بالمجال والعلاقات بين المفاهيم التي تمثلها هذه المصطلحات من مجموعة نصية مكتوبة بلغة طبيعية، وتتميزها بلغة الأنطولوجيا لسهولة استرجاعهم.

تتضمن عملية توليد الأنطولوجيا: استخلاص المصطلحات والمفاهيم أو العبارات من النصوص باستخدام طرائق المعالجة اللغوية مثل علامات نوع الكلام part of speech tagging وتقطيع العبارات phrase chunking، واستخلاص العلاقات باستخدام تقنيات إحصائية ورمزية.

وقد جاءت ضرورة تعلّم الأنطولوجيا من عوامل عدة: منها كلفة بناء الأنطولوجيا المرتفعة، والتغير المستمر في العلوم والمعارف.

استخدمت في العقود الماضية تقنيات عدة من أجل توليد الأنطولوجيا (Asim et al., 2018) منها: معالجة اللغات الطبيعية، والتعلم الآلي واسترجاع المعلومات والتنقيب عن المعلومات. وفرت تقنيات التنقيب عن المعلومات واسترجاع المعلومات والتعلم الآلي آلية إحصائية لاستخلاص المصطلحات والمفاهيم وارتباطاتهم. ومن جانب آخر لعبت تقنيات معالجة اللغات الطبيعية دوراً في معظم مراحل تعلم الأنطولوجيا

كما استخدمت تقنيات التقابل Mapping أو الدمج merging أو التكامل integration أو alignment في العديد من التطبيقات مثل: ((Szymanski et al., 2012)، (Fernando et al., 2009)، (Soor, 2015)، (John et al., 2017)، (Juzaddin et al., 2015)، (Ponzetto et al., 2010) من أجل: دمج عدة أنطولوجيات، أو توسيع الأنطولوجيا بتقابل مفاهيمها مع مصادر معرفية شبه مهيكلة، مثل: الموسوعات، وذلك إما بربط عناوين أو اختصاصات، أو خصائص مقالات الموسوعات بالمفاهيم في شبكة الكلمات.

لا توجد قياسات كثيرة ضابطة لتقييم تعلّم الأنطولوجيا لأنّ علم تعلّم الأنطولوجيا حديث نوعاً ما. كما أن نتائج هذا النوع من التعلّم ليست قائمة بسيطة بتصنيفات ثنائية وإنما لها بنية معقدة. فكما صرح الباحث في (Smith, 2004) أن هناك احتمالات متعددة للتجريد في مجال واحد تختلف باختلاف استخدامها واختلاف مجموعات

الأشخاص. فإذا لم تقارن نتائج الخوارزمية المطبقة في تعلّم الأنطولوجيا مع أنطولوجيا جيدة مبنية يدوياً فلا يمكن تقييم جودتها. ولأن الأنطولوجيا ليست بحد ذاتها موضع الاهتمام وإنما التطبيقات التي تستخدم فيها كمصادر لتحسين الأداء، فهناك طرائق عدة يمكن تطبيقها للتقييم الآلي للأنطولوجيات (Biemann, 2005) تتضمن: التقييم باستخدام معيار الذهب a Gold Standard وذلك بمقارنة الأنطولوجيا المتولدة مع أنطولوجيا المرجعية، التقييم المعتمد على التطبيق Application-based وذلك بقياس مدى مساعدة الأنطولوجيا المتولدة في تحسين نتائج التطبيق، والتقييم المساق بالبيانات Data driven.

5-4 شبكة الكلمات العربية:

شبكة الكلمات العربية Arabic wordnet هي مستودع معرفي هرمي البنية غني بالعلاقات المعجمية الصريحة أنشئ من قبل خبراء لغويين. وهو إنتاج مشروع بحثي طوّر في جامعة Princeton للحصول على معارف معجمية للغة الأم وحفظه في قاموس يمكن قراءته من قبل الآلة (Fellbaum et al., 1998). تحتوي كلمات مصنّفة إلى أسماء وأفعال وصفات وظروف، وكلمات أخرى من نوع ضمائر وحروف جر وعطف. تُنظّم الكلمات ضمن شبكة الكلمات في مجموعات مترادف (المفاهيم) تسمى Synsets، وهي تشكل البنية الأساسية للشبكة. يمكن لكلمة واحدة أن تحمل أكثر من معنى، يعطى لكل معنى فيها عنصر يسمى معنى الكلمة word sense. تحتوي مجموعة المرادفات على معاني للمساعدة في وصفها. لكل مجموعة مرادفات شرح gloss يوصف معنى الكلمة. كما ترتبط معظم المجموعات بأمثلة استخدام.

ترتبط المعاني والمفاهيم بعلاقات متنوعة. تسمى العلاقات التي تربط بين المعاني العلاقات الدلالية Semantic Relations، بينما تسمى تلك التي تربط المفاهيم علاقات معجمية Lexical Relations. ترتبط الأسماء بعلاقات دلالية مثل: الجزء/التضمين Hyponym/Hypernym، والانتماء/الاشتغال Meronym/Holonym، وجزء من/لها جزء Meronym/Holonym، ومستوى/له مستوى Meronym/Holonym. تشكل هذه العلاقات تصنيفات ذات بنية هرمية. تنتمي بعض المجموعات إلى تصنيف

واحد بينما ينتمي بعضها الآخر إلى أكثر من تصنيف. كما يوجد علاقات معجمية أخرى مثل المضاد antonyms، والمقاد derived-from. تكون العلاقات المعجمية بين معاني الكلمات وليس بالضرورة بين المفاهيم.

تُمثل كلمات شبكة الكلمات بتنسيق خاص ولكل كلمة نوع الكلام (POS) Part-of-Speech ورقم المعنى. وتوجد أربعة أنواع للكلام في الشبكة هي: n للأسماء، و v للأفعال، و a للصفات، و r للضمائر.

وقد تُرجمت شبكة الكلمات إلى لغات عدة من ضمنها العربية فشكّلت شبكة الكلمات العربية Arabic WordNet AWN وهي أنطولوجيا عربية معجمية استخدمت على نحو واسع، إلا أنها تقتصر إلى بعض المفاهيم والعلاقات العربية ولها تحديات عدة (Black et al., 2006) لأنها تشكّلت عن طريق الترجمة.

5-5 الموسوعة العربية

تعد الموسوعة العربية Arabic Encyclopedia أكبر الموسوعات العربية المتوفرة ورقياً وإلكترونياً، مكتوبة بأقلام عربية بمصطلحات وتعابير عربية، وهي ليست ترجمة لثقافة ومصطلحات غريبة عن اللغة العربية. وهي من تأليف باحثين عرب، ألفت في هيئة الموسوعة العربية في الجمهورية العربية السورية، صدرت نسختها الإلكترونية على موقع هيئة الموسوعة العربية منذ عام 2010 ومتاحة مجاناً للعموم. تتألف من حوالي 15000 مقالة في مجالات المعرفة الإنسانية جميعها التي تتضمن: الحضارة العربية، والعلوم الإنسانية، والآداب واللغات، والعلوم الطبية، والعلوم البحتة، والعلوم التطبيقية، والعلوم القانونية، والتربية والفنون.

تتألف كل مقالة في الموسوعة من نص يصف موضوع المقالة مع روابط للمواضيع ذات الصلة. الهدف من الروابط تزويد القراء بمعلومات إضافية مرتبطة بالموضوع. إضافة للنص الذي يصف المقالة فهناك مجموعة مصطلحات تُشكّل عنواناً فريداً للمقالة. وعندما تُناقش مقالتان أو أكثر موضوعاً بكلمة تحمل أكثر من معنى يرفق العنوان عادة بمصطلح وصفي للتمييز بين المادتين.

توصف المقالات الموجودة في الموسوعة العربية، كما في الموسوعات الأخرى مواضيع عدة. يمكن عدّ كل مقالة مفهوماً يرتبط بمحتوى نص المقالة الذي يصف موضوعها الأساسي.

5-6 الأبحاث والدراسات السابقة لتوليد أنطولوجيا عربية

يمكن تلخيص أهم المشاريع والدراسات السابقة لتوليد أنطولوجيا عربية بمايلي:

اعتمد في توليد أنطولوجيا عربية على طريقتين: الأولى: ترجمة أنطولوجيات أجنبية، كما في مشروع Arabic WordNet (AlBarhamtoshy et al., 2009) حيث قام الباحثون فيه بتصميم وتنفيذ شبكة كلمات عربية عن طريق ترجمة شبكة الكلمات الإنكليزية WordNet إلى العربية باستخدام قاموس متعدد اللغات. كما وُلدت شبكة الكلمات العربية في مشروع (Rodríguez et al., 2008) بتوسيع شبكة الكلمات العربية بشكل شبه آلي باستخدام قواعد صرفية وقواعدية وتطبيق استدلال بايزن على شبكة الكلمات الإنكليزية.

والطريقة الثانية: توسيع أنطولوجيات عربية مترجمة اعتماداً على مصادر معرفية مترجمة، كما في مشروع (Al khalifa et al., 2008) حيث وُسعت شبكة الكلمات العربية آلياً بإضافة أسماء كيانات عربية من عناوين مقالات موسوعة الويكيبيديا، وفي مشروع (Juzaddin et al., 2015) حيث وسعت شبكة الكلمات العربية بالتقابل بين مرادفات شبكة الكلمات العربية وخصائص مقالات موسوعة الويكيبيديا المترجمة إلى العربية. واستخدمت الأنطولوجيا اللغوية Gold كنواة في مشروع أنطولوجيا الخليل (Aliane et al., 2009) لبناء البنية التحتية لأنطولوجيا عربية مركزية. كما قام الباحثون في مشروع (Mazari et al., 2012) ببناء أنطولوجيا في مجالات لغوية تخص اللغة العربية معتمدين المفاهيم العامة في أنطولوجيا Gold ومجموعة خصائص نصوص عربية واستخدام تقنيات إحصائية تعتمد على توارد العبارات في تلك النصوص، كما اعتمد الترافق لربط المفاهيم المستخلصة بالأنطولوجيا بعلاقات هرمية وغير هرمية.

وطور الباحثون مشروع الأنطولوجيا العربية بجامعة بيرزيت ((Jarrar et al., 2011)، (Jarrar et al., 2021)) أنطولوجيا عربية بتوسيع شبكة الكلمات العربية اعتماداً على المعاجم باستخراج المفاهيم العربية ومقابلتها مع شبكة الكلمات العربية.

كما استخدمت طرائق إحصائية لإنشاء أنطولوجيا عربية متخصصة بمجال معين، كما في مشروع (Alsudias et al., 2020) حيث وُلدت أنطولوجيا عربية طبية متخصصة بالأمراض المعدية حيث استخرجت

المصطلحات من مستندات طبية في هذا المجال آلياً وجرى مكاملتها مع معاجم علمية. وفي مشروع (Mohammed et al., 2014) حيث وُلدت علاقات دلالية آلياً من نصوص عربية باعتماد الأنماط القواعدية. كما وُلدت في مشروع (Belkredim et al., 2008) أنطولوجيا الديار DEAR-ONTO وهي أنطولوجيا عربية اشتقاقية بينية، استُخدمت فيها الأفعال السماعية كجذور وحددت الأصناف المختلفة المشكلة للكلمات المشتقة منها. كما كانت هناك أبحاث عدة ومشاريع لتوليد أنطولوجيات للقرآن الكريم مثل شبكة الكلمات العربية القرآنية Quranic Arabic WordNet (QAWN) في (AlMaayah et al., 2016) وذلك بتوليد المرادفات آلياً مستندين على قواميس عربية. وأنطولوجيا القصص القرآنية باستخدام تقنية التقابل في مشروع (Al-Salhi et al., 2022). وفي مشروع أنطولوجيا الأزهرى Azhary (Ishkewy et al., 2014) وأوجدت أنطولوجيا عربية معجمية وذلك باستخراج الكلمات والمفاهيم العربية آلياً من نصوص القرآن الكريم، وربطها يدوياً بعلاقات، استناداً إلى معاجم عربية عربية.

5-7 توليد الأنطولوجيا العربية بالتقابل بين شبكة الكلمات العربية والموسوعة

العربية

نتيجة دراسة المشاريع السابقة في مجال توليد أنطولوجيا عربية تبين لنا أنه لا يوجد أنطولوجيا عربية تضم المفاهيم العربية ومجموعة ترادفات المصطلحات العربية وعلاقاتها الدلالية، إلا في بعض المشاريع البحثية غير المكتملة والتي تعتمد في توليدها على الترجمة سواء من مصادر معرفية مثل موسوعة الويكيبيديا أم من أنطولوجيا مثل شبكة الكلمات الإنكليزية. لذا حاولنا في هذا البحث توليد أنطولوجيا عربية اعتماداً على توسيع شبكة الكلمات العربية وإغنائها بالاستعانة بخصائص الموسوعة العربية التي تعدّ أفضل مصدر معرفي عربي غني بالمعلومات والمفاهيم العربية في المجالات كافة وذلك باستخدام تقنية التقابل.

قسمت طريقة توليد الأنطولوجيا العربية إلى مراحل عدة كما هو موضح في الشكل رقم (2-4) وهي:

توليد خصائص الموسوعة العربية، تجهيز بيانات شبكة الكلمات العربية، تعلم الأنطولوجيا بالتقابل بين

الموسوعة العربية وشبكة الكلمات. ونقدم فيما يلي شرحاً توضيحياً لتلك المراحل.

5-7-1 توليد خصائص الموسوعة العربية:

الهدف من توليد خصائص الموسوعة (Fawcett T. E. 1993) إثراء تمثيل المستند النصي من خلال تعزيزه بخصائص مُستخرجة منها. تتضمن هذه المرحلة قسمين، قسم لاستخراج الخصائص وتثقيلها، وقسم لحساب العلاقات بين المصطلحات.

5-7-1-1 استخراج الخصائص وتثقيلها:

تتضمن هذه المرحلة المعالجة الأولية لنصوص مقالات الموسوعة وفهرستها فهرسة آلية لاستخراج المصطلحات وتثقيلها اعتماداً على تواريخها في نصوص الموسوعة وفقاً للخطوات المبينة في الشكل (5-1) وهي: التقطيع Tokenization، والتقييس Normalization، وإزالة كلمات التوقف Removing Stopword، والتجريد Stemming، وتثقيل الكلمات المفتاحية Weighing of keywords.



الشكل (5-1) خطوات المعالجة الأولية للنصوص العربية

التقطيع: استخدمت طريقة التقطيع المطبقة في أداة AraNLP (Althobaiti et al., 2014)، وذلك بتحويل الشريط المحرفي (المستند) إلى مجموعة مصطلحات، اعتماداً على محددات مثل المسافة البيضاء والفاصلة المنقوطة والاقتباسات.

التقييس: استُخدمت طريقة التقييس المطبقة في أداة AraNLP (Althobaiti et al., 2014): وذلك بإزالة الأحرف غير العربية، مثل: الرموز وعلامات الترقيم الخاصة بإزالة علامات التشكيل، واستبدال "ا" بـ "أ" ، "آ" ، و"ي" بـ "ى" ، و"هـ" بـ "ة" ، و"و" بـ "ؤ" ، و"ي" بـ "ئ" وإزالة التطويل.

إزالة كلمات التوقف: تحتوي اللغة العربية مثل باقي اللغات على كلمات وظيفية لاتضيف معنى في استرجاع المعلومات، فيمكن حذفها من الجمل دون التأثير على محتواها الأساسي. اعتمدت قائمة التوقف العامة المستخدمة

في أداة AraNLP (Althobaiti et al., 2014)، وهي القائمة المعتمدة في معظم التطبيقات مع بعض التعديلات عليها.

التجريد: التجريد بالعربية عملية شاقة ودقيقة، لأنها صرفية واشتقاقية، لذا اعتمدت طريقة تجريد بحذف السوابق واللواحق المطبقة في مجرد خوجة (Khoja et.al., 1999) المستخدم على نحو واسع في مجال تجريد الكلمات العربية، وذلك بإزالة السوابق مثل (ال، وال، بال، كال، لل، ل، ف، ب، ك) واللواحق مثل (كن، هما، كما، ها، ك، ه)، ومقارنة الكلمات المجردة مع قائمة جذور عربية.

تثقيل الكلمات المفتاحية: تثقل المصطلحات (الكلمات المفتاحية) الواردة في مستند معين لمعرفة أهميتها فيه، استخدمت المعادلة (1-5) وهي علاقة التثقيل المشهورة (Term frequency – Inverse TF-IDF) (Salton, 1989) document frequency، من أجل حساب وزن المصطلحات والتي قيس بـ '0.1' وجمعه الذي تم اختياره تجريبياً من أجل تقليل التفاوت الكبير بين القيم الذي يختلف حسب طول المستند.

$$w = 0.1 + 0.1 \left(\frac{t_{fi}}{\max_{t_f}} \right) \log \left(\frac{N}{n_i} \right) \quad (1 - 5)$$

حيث:

- t_{fi} : تردد المصطلح في المستند.
- Max_{tf} : أعلى تردد مصطلح في المستند.
- N : عدد الجمل في المستند.
- n_i : عدد المستندات التي تحتوي المصطلح i .

5-7-1-2 حساب العلاقات بين الخصائص :

استخدمت تقنيات إحصائية اعتماداً على تردد حدوث المصطلحات في المجموعة، لاكتشاف العلاقات بين المصطلحات والمفاهيم المستخرجة من الموسوعة العربية وحسابها، والعلاقات التي تولدت منها هي علاقة الترافق Co-Occurance Relations، وعلاقة التشابه Similarity Relations.

علاقة الترافق: حسبت هذه العلاقة (باستخدام الإجرائية co_occur المبينة في الملحق B)، اعتماداً على

ترافق مصطلحين معاً، وذلك باستخدام المعادلة (2-5) (Van Rijsbergen, 1976)، والتي تربط بين مصطلحين اعتماداً على تواردتهما المشترك.

$$CoOccurance(t_u, t_v) = \frac{CoOccurance\ Frequency}{\sqrt{Frequency(t_u) * Frequency(t_v)}} \quad (2-5)$$

حيث:

- t_u و t_v : المصطلحان المرتبطان بعلاقة ترافق.
- CoOccuranceFrequency: هو تردد حدوث المصطلحين t_u و t_v معاً في مجموعة مستندات الموسوعة.
- frequency(t_u): هو تردد حدوث المصطلح t_u في المجموعة.
- frequency(t_v): هو تردد حدوث المصطلح t_v في المجموعة.

علاقة التشابه: ترتبط المصطلحات مع بعضها بعضاً باعتماد تواردتها في الموسوعة، حسبت هذه العلاقة

باستخدام علاقة التجيب بين أشعة المصطلحات (باستخدام الإجرائية Similar المبينة في الملحق B)، وذلك بتمثيل كل مصطلح بشعاع مركباته أوزان هذا المصطلح في مستندات الموسوعة، وفقاً للمعادلة (3-5) Gomaa et al.,. (2013).

$$Sim_{t_u, t_v} = \frac{\sum_{i=1}^p w_{ui} w_{vi}}{\sqrt{\sum_{i=1}^p w_{ui}^2} \sqrt{\sum_{i=1}^p w_{vi}^2}} \quad (3-5)$$

حيث:

P: مجموعة المستندات في الموسوعة العربية.

w_{ui} : وزن المصطلح u في المستند i .

w_{vi} : وزن المصطلح v في المستند i .

5-7-2 تجهيز بيانات شبكة الكلمات العربية:

استخدمت أداة تجهيز بيانات شبكة الكلمات العربية لفهرسة مفاهيم ومصطلحات الشبكة التي تتكون من 11270 مفهوماً و 23496 مصطلحاً. وذلك بتطبيق طريقة فهرسة مصطلحات الموسوعة العربية نفسها، من أجل إتاحة إمكانية مقارنتها مع مفاهيم ومصطلحات الموسوعة العربية في عملية التقابل المقترحة، وفق المراحل التالية: التقييس، وإزالة كلمات التوقف، والتجريد. والموضحة بالشكل (5-2).

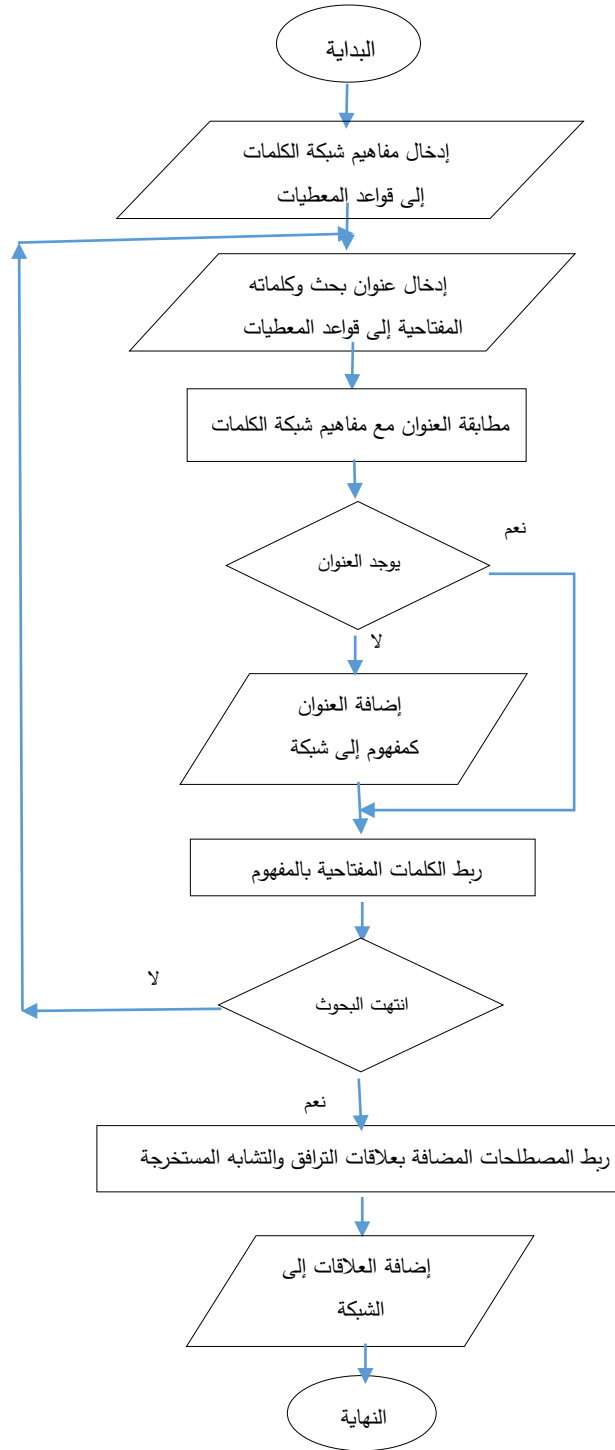


الشكل (5-2) مراحل تجهيز بيانات شبكة الكلمات العربية

5-7-3 تعلم الأنطولوجيا بالتقابل بين خصائص الموسوعة وشبكة الكلمات:

وُلدت الأنطولوجيا العربية اعتماداً على خوارزمية تعلم الأنطولوجيا الموضحة في الشكل (5-3). استند في هذه الخوارزمية إلى تقنية التقابل بتوسيع أنطولوجيا شبكة الكلمات العربية اعتماداً على البيانات الناتجة عن مرحلة استخراج خصائص الموسوعة العربية (باستخدام إجرائية التقابل Mapping المبينة في الملحق B). وفقاً للخطوات التالية:

- 1- مطابقة مفاهيم شبكة الكلمات مع عناوين مقالات الموسوعة، وإضافة عناوين المقالات غير الموجودة في الشبكة كمفاهيم جديدة باستثناء أسماء الأعلام.
- 2- إضافة المصطلحات المهمة في مستندات الموسوعة (التي لها ثقل يساوي 0.7 أو يتجاوز في مستندات الموسوعة، وقد اختير هذا الرقم تجريبياً من أجل تحسين جودة النتائج المستخرجة آلياً) إلى الشبكة بضمها إلى المفهوم المناسب في الشبكة.
- 3- ربط المصطلحات والمفاهيم المضافة إلى الشبكة بالعلاقات المستخرجة، وهي علاقات الترافق والتشابه (التي لها ثقل يتجاوز 0.5، وقد اختير هذا الرقم تجريبياً من أجل تحسين جودة النتائج المستخرجة آلياً).



الشكل (3-5) خوارزمية تعلم الأنطولوجيا

5-7-4 تقييم الأنطولوجيا العربية

اعتمد في تقييم الأنطولوجيا العربية المولدة على التطبيق، وذلك من خلال اعتماد الأنطولوجيا المتولدة في نظام التلخيص المقترح في هذا البحث، حيث قورنت نتائج تلخيص نظام التلخيص باعتماد الأنطولوجيا مع نتائج تلخيص نظام التلخيص غير المعتمد على الأنطولوجيا، وذلك من خلال تطبيق النظامين لتلخيص مجموعة مستندات معيارية هي مجموعة الملخصات العربية من (Essex (EASC (El-Haj et al., 2010)، وقياس متوسطات مقاييس الدقة والاستدكار ومعامل F في ملخصات كلا النظامين . فكانت نتائج التلخيص باستخدام الأنطولوجيا أفضل كما هو موضح في الفصل السابع.

لذا يمكن أن تُعدّ هذه الأنطولوجيا نواة لبناء أنطولوجيا عربية غير مترجمة يمكن اعتمادها في تطبيقات الويب الدلالي وغيرها من التطبيقات الدلالية التي تعالج اللغة العربية، ويمكن إغناؤها وتوسيعها للحصول على أنطولوجيا عربية أشمل وأدق.

الفصل السادس

توليد الملخص الآلي للنصوص العربية اعتماداً على الأنطولوجيا العربية

Generation of Automatic Arabic Summarizer Using Arabic Ontology

1-6 مقدمة:

تضمنت الفصول السابقة شرحاً ملخصاً عن أهم طرائق ومشاريع التلخيص الآلي للنصوص باللغات العالمية وأحدثها عموماً، وسلط الضوء على دراسات التلخيص الآلي للنصوص العربية خصوصاً ومشاريعه. وطريقة توليد أنطولوجيا عربية لاعتمادها في التلخيص الآلي. يتضمن هذا الفصل كيفية استخدام الأنطولوجيا العربية المتولدة من أجل تلخيص النصوص العربية، وتفاصيل تصميم نظام التلخيص وتطبيقه وتقييمه.

2-6 مراحل النظام

اقترح نظامان لتلخيص النصوص العربية وهما نظام تلخيص باعتماد الأنطولوجيا ونظام تلخيص إحصائي لايعتمد الأنطولوجيا، واستخدم الأخير من أجل المقارنة وتقييم النظام المعتمد على الأنطولوجيا. وقد اعتمدت الأنطولوجيا العربية المولدة بالتقابل بين شبكة الكلمات العربية وخصائص الموسوعة العربية. وفيما يلي نقدم شرحاً عن نظامي التلخيص المقترحين.

يطبق النظامان المقترحان تلخيصاً استخراجياً عاماً أحادي المستند وأحادي اللغة، ويتشاركان في مراحل التلخيص الأساسية الموضحة في الشكل (3-4) مع اختلافات في تنفيذها توضح أثناء الشرح، وهذه المراحل هي: المعالجة الأولية Text Preprocessing لمستندات المصدر، والتحليل واستخراج الخصائص وصياغتها Analysis and Feature extraction and formulation، وتقييم الجمل وتوليد الملخص Sentence evaluation and summary generation. وفيما يلي نقدم شرحاً توضيحياً عن طريقة تنفيذ هذه المراحل.

3-6 المعالجة الأولية للوثيقة

هذه المرحلة هي المرحلة الأولية في جميع طرائق التلخيص تقريباً. والغرض الرئيسي منها هو إعداد مستند الدخل النصي للمعالجة في المراحل الأخرى. يتضمن نظام تلخيص النصوص المقترح العمليات المتسلسلة التالية للمعالجة الأولية: التقطيع Tokenization، والتنقيس Normalization، وإزالة كلمات التوقف StopWord Removal، والتجريد Stemming، (Abdelkrime et al., 2015)، (Litvak et al., 2016)، (Thomas et al., 2015)). يوضح الشكل (1-6) خطوات المعالجة الأولية في تلخيص النصوص.



الشكل (1-6) خطوات المعالجة الأولية في تلخيص النصوص

1-3-6 التقطيع:

تبدأ المعالجة الأولية للنص بعملية التقطيع Tokenization استناداً إلى تحليل صرفي يعتمد على علامات الترقيم لتقسيم المستند إلى جمل بناءً على النقطة (.) وعلامة الاستفهام (?) وعلامة التعجب (!) والفواصل (،) والفاصلة المنقوطة كـمحددات. ثم تُقسم هذه الجمل إلى مقاطع استناداً إلى محددات مثل المسافة البيضاء والفاصلة المنقوطة والاقتباسات.

2-3-6 التنقيس:

قد تظهر في اللغة العربية بعض الحروف العربية بأشكال مختلفة، وتستخدم حروف بدلاً من حروف لأن أشكالها متشابهة، كما يستخدم الكتاب علامات التشكيل في نصوصهم، التي يمكن أن تؤدي إلى اختلافات للمصطلح نفسه، ومن ثم تؤثر على حساب بعض الخصائص مثل التردد (TF). لذلك فإن عملية التنقيس مطلوبة لتوحيد الأشكال المختلفة للحرف نفسه لتجنب مثل هذه الاختلافات. استُخدمت طريقة التنقيس المطبقة في أداة (Althobaiti et al., 2014) AraNLP.

6-3-3 إزالة كلمات التوقف:

كلمات التوقف Stop-word (مثل الضمائر وحروف الجر وحروف العطف وما إلى ذلك) هي كلمات غير مهمة تظهر كثيراً في المستندات لتكوين الجمل (Kanan et al., 2004). نظراً لأن هذه الكلمات ليست مفيدة (لا تضيف معلومات)، فيمكن حذفها من الجمل دون التأثير على المحتوى الأساسي للجملة. في الواقع، هناك عديد من طرائق قائمة التوقف المستخدمة لإزالة كلمات التوقف من النص بما في ذلك، قائمة توقف عامة، قائمة توقف بمجال معين، وقائمة توقف مدمجة. اعتمد في هذا البحث على قائمة التوقف العامة المستخدمة في أداة AraNLP، وهي القائمة المعتمدة في معظم التطبيقات مع بعض الإضافات.

6-3-4 التجريد:

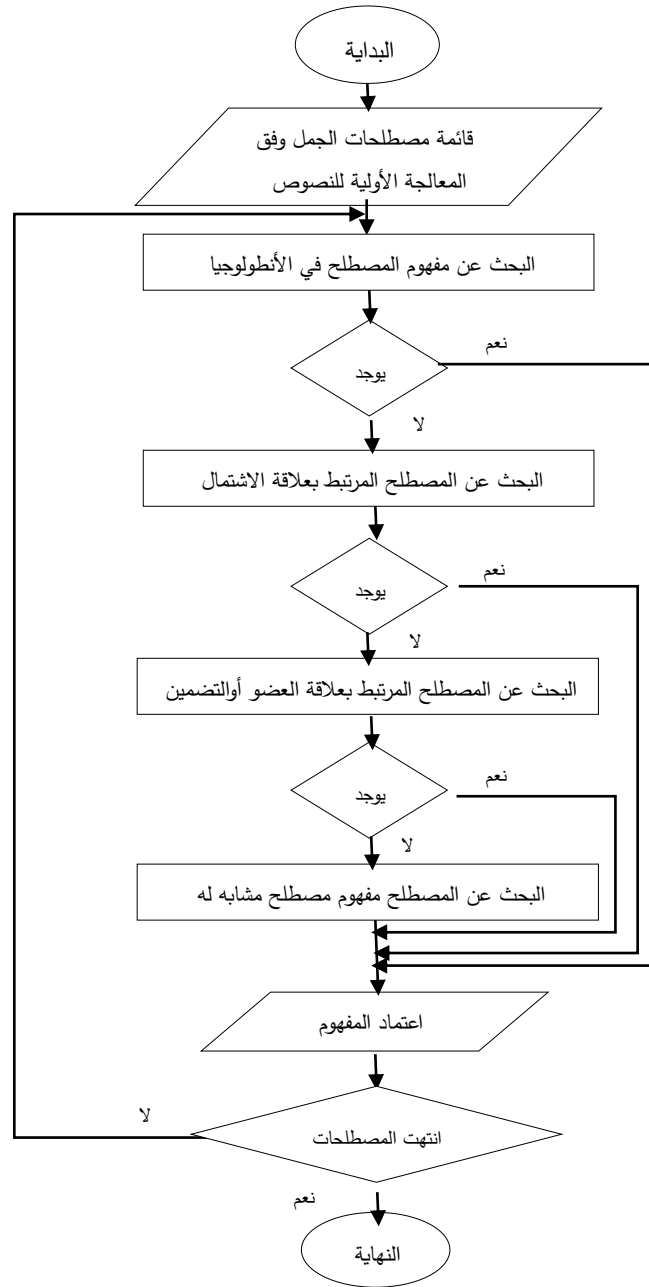
التجريد Stemming بالعربية عملية شاقة ودقيقة، لأن اللغة العربية هي لغة صرفية واشتقاقية بحتة، مما يعني أن الكلمات العربية يمكن أن يكون لها عديد من الأشكال المختلفة ولكنها تشترك في المعنى المجرد نفسه. إن عملية التجريد هي عملية إزالة بعض اللواحق (مثل البادئات، واللواحق، واللواحق اللاحقة) من الكلمة وذلك بتحويل الأشكال والمشتقات المختلفة للكلمة إلى شكل موحد واحد مثل الجذر أو الجذع. في اللغة العربية، هناك طريقتان شائعتان للتجريد، التجريد المعتمد على الجذور والتجريد الخفيف (Mustafa et al., 2017). قورنت هذه التقنيات في العمل المقدم في (Alami et al., 2016) فيما يتعلق بتلخيص النصوص باستخدام اثنين من المجردات العربية المعروفة. وأظهرت تجاربهم أنه في تلخيص النص العربي يفضل تجريد الجذر على التجريد الخفيف. بناءً على هذه النتائج استخدمت طريقة مجرد خوجة (khoja, 1999) وقواعد بياناتها.

6-4 التحليل واستخراج الخصائص وصياغتها:

يجري في مرحلة التحليل واستخراج الخصائص وصياغتها Analysis and Feature Extraction and Formulation اختيار الجمل ذات الأهمية أوالتي لها المواءمة العليا استناداً على توظيف مجموعة خصائص لتوليد ملخص متماسك تحوي الأفكار الرئيسة في مستند معين. اعتمد في اختيار الخصائص على فرضيات وتجارب

وبعض الدراسات السابقة ((Ferreira et al., 2012)، (Meena et al., 2014)، (Meena et al., 2016)) وما يناسب اللغة العربية. لذا اعتمدت الخصائص التالية: خاصية العبارات المفتاحية، خاصية التشابه مع العنوان، خاصية موقع الجملة، طول الجملة، العبارات البارزة، العبارات غير هامة. وقد حُسبت بعض خصائص التلخيص بالاعتماد على طريقتين، طريقة تخص التلخيص الإحصائي المقترح استناداً إلى توارد المصطلحات في الجمل والمستند وتكرارها، وطريقة تخص التلخيص بالاعتماد على الأنطولوجيا استناداً إلى توارد المفاهيم في الجمل والمستند وتكرارها، وذلك بعد إعادة المصطلحات إلى مفاهيمها، باستخدام العلاقات الدلالية في الأنطولوجيا العربية وفقاً لخوارزمية إعادة مصطلحات الجمل إلى مفاهيمها المبينة في الشكل (6-2) والتي تتضمن الخطوات الآتية:

1. يسند إلى المصطلح المفهوم المناسب له حسب علاقة الترادف synset في الأنطولوجيا.
 2. يسند إلى مفهوم المصطلح أحد المصطلحات المرتبطة به وفق علاقات الاشتمال إذا لم تكن له علاقة ترادف حسب البند 1.
 3. يسند إلى مفهوم المصطلح أحد المصطلحات المرتبطة به وفق علاقات العضو إذا لم تكن له علاقة اشتمال حسب البند 2.
 4. يسند إلى مفهوم المصطلح أحد المصطلحات المرتبطة به وفق علاقات التضمن إذا لم تكن له علاقة عضو حسب البند 3.
 5. يسند إلى مفهوم المصطلح أحد مفاهيم المصطلحات المشابهة له، والتي ترتبط معه بعلاقات تشابه، إذا لم يرتبط المصطلح بعلاقات حسب البنود السابقة.
 6. يعد المصطلح مفهوماً إذا لم يرتبط المصطلح بأي من العلاقات وفق البنود السابقة.
- وفيما يلي نقدم شرحاً تفصيلياً لتلك الخصائص المستخدمة وطريقة حسابها في نظامي التلخيص المقترحين.



الشكل (2-6) خوارزمية إعادة مصطلحات الجمل إلى مفاهيمها

6-4-1 خاصية العبارات المفتاحية:

إن إمكانية وجود فكرة أساسية في الجملة مشروطة باحتوائها على عبارة أو كلمات مفتاحية من شأنه زيادة

أهميتها بالنسبة للجملة الأخرى ((El-Shishtawy et al., 2012)، ((El-Beltagy et al., 2009)).

اعتمد في هذا البحث إجرائية استخراج العبارات المفتاحية اعتماداً على ثقل وحدات في الجمل بطريقتين، طريقة تخص التلخيص الإحصائي المقترح، وطريقة تخص التلخيص باعتماد الأنطولوجيا المقترح. نوضح فيما يلي طريقة حساب هذه الخاصية في النظامين.

6-4-1-1 خاصية العبارات المفتاحية في التلخيص الإحصائي:

تحتسب درجة خاصية العبارة المفتاحية في نظام التلخيص الإحصائي المقترح: باستخدام إجرائية بناءً على ثلاثة من أهم العوامل وهي: تردد المصطلح في الجملة، عدد الجمل في المستند، وعدد الجمل التي تحتوي المصطلح في المستند. استخدمت المعادلة (1-6) وهي علاقة التثقل المشهورة وهي علاقة التثقل المشهورة (Term TF-IDF frequency – Inverse document frequency) (Salton,1989) ، من أجل حساب وزن العبارات المفتاحية في جملة معينة، والتي تُقيس بضرب العدد '0.1' وجمعه، والذي اختير تجريبياً من أجل تقليل التفاوت الكبير بين القيم المختلفة حسب طول الجملة.

$$S_{key} = 0.1 + 0.1 \left(\frac{t_{fi}}{\max_{t_{fi}}} \right) \log\left(\frac{N}{n_i}\right) \quad (1-6)$$

حيث:

- t_{fi} : تردد المصطلح في الجملة.
- $\max_{t_{fi}}$: أعلى تردد في المستند.
- N : عدد الجمل في المستند.
- n_i : عدد الجمل التي تحتوي المصطلح i .

6-4-1-2 خاصية العبارات المفتاحية في التلخيص باعتماد الأنطولوجيا:

استند حساب درجة خاصية العبارات المفتاحية في التلخيص باعتماد الأنطولوجيا على: حساب الثقل استناداً إلى المفاهيم بدلاً من المصطلحات، حيث يعاد المصطلح إلى مفهومه الدلالي باستخدام الأنطولوجيا العربية المقترحة (وفقاً لخوارزمية إعادة مصطلحات الجمل إلى مفاهيمها المبينة في الشكل (3-6))، ومن ثم تُنقل المفاهيم في

الجمال استناداً إلى تواردها مفاهيمها في الجملة والمستند وفقاً لعلاقة التثقيل وفقاً للمعادلة (2-6) المشابهة للعلاقة المطبقة في التلخيص الإحصائي.

$$S_{key} = 0.1 + 0.1 \left(\frac{C_{fi}}{\max C_{fi}} \right) \log \left(\frac{N}{n_i} \right) \quad (2-6)$$

حيث:

- C_{fi} : تردد المفهوم في الجملة.
- N : عدد الجمل في المستند.
- n_i : عدد الجمل التي تحتوي المفهوم i .

2-4-6 خاصية التشابه مع العنوان:

عُرفت خاصية التشابه مع العنوان Similarity With Title على أنها التشابه أو التداخل بين جملة معينة وعنوان المستند ((Fattah et al., 2009)، ((Nobata et al., 2009)).

عُدَّت في هذا البحث الجملتان الأوليتان في المستند هما عنوان المستند لأن هناك مستندات دون عنوان. وحُسبت خاصية التشابه مع العنوان بطريقتين، طريقة تخص التلخيص الإحصائي المقترح، وطريقة تخص التلخيص باعتماد الأنطولوجيا المقترح. نوضح فيما يلي طريقة حساب هذه الخاصية في النظامين.

1-2-4-6 خاصية التشابه مع العنوان في التلخيص الإحصائي:

تُحسب درجة خاصية التشابه مع العنوان في التلخيص الإحصائي المقترح وفقاً للمعادلة (3-6) ((Fattah et al., 2009)، ((Abuobieda et al., 2012)) باعتماد ثقل المصطلحات الواردة في الجملة والعنوان وتكرارها.

$$S_{TitleSim} = Similarity(Title, S_i) * \sqrt{1 + (KP_T \cap KP_{Si})} \quad (3-6)$$

حيث:

- S_i هي الجملة الحالية.
- KP_T عدد العبارات المفتاحية الموجودة في العنوان.
- KP_{Si} عدد العبارات المفتاحية الموجودة في الجملة S_i .

- $(KP_T \cap KP_{Si})$ هو عدد العبارات المفتاحية المشتركة بين الجملة Si والعنوان.
- $Similarity(Title, Si)$ هي درجة تشابه العنوان مع جملة Si التي تحسب بمقياس تشابه التجيب باستخدام العلاقة المشهورة لحساب تشابه النصوص (Gomaa et al., 2013) وفقاً للمعادلة رقم (4-6)

$$Similarity_{Title, Si} = \frac{\sum_{t=1}^n w_{tSi} w_{tTitle}}{\sqrt{\sum_{t=1}^n w_{tSi}^2} \sqrt{\sum_{t=1}^n w_{tTitle}^2}} \quad (4-6)$$

حيث:

- $Similarity_{Title, Si}$ التشابه بين الجملة Si والعنوان $Title$,
- n عدد المصطلحات الفريدة في المستند D .
- w_{tSi} وزن المصطلح t في الجملة Si .
- w_{tTitle} وزن المصطلح t في العنوان $Title$.

2-2-4-6 خاصية التشابه مع العنوان في التلخيص باعتماد الأنطولوجيا:

تُحسب درجة خاصية التشابه مع العنوان في التلخيص باعتماد الأنطولوجيا اعتماداً على المفاهيم، وذلك بإعادة المصطلحات إلى مفاهيمها (وفقاً لخوارزمية إعادة مصطلحات الجمل إلى مفاهيمها المبينة في الشكل رقم (2-6))، ومن ثم تحسب درجة خاصية التشابه مع العنوان باستخدام علاقة مشابهة للعلاقة المشابهة للعلاقة المستخدمة في التلخيص الإحصائي، مع اختلاف اعتمادها على ثقل المفاهيم الواردة في الجملة والعنوان وتكرارها بدلاً من المصطلحات، وفقاً للمعادلة (5-6).

$$S_{TitleSim} = Similarity(Title, Si) * \sqrt{1 + (KP_c \cap KP_{Si})} \quad (5-6)$$

حيث:

- Si هي الجملة الحالية.
- KP_c عدد المفاهيم المفتاحية الموجودة في العنوان.
- KP_{Si} عدد المفاهيم المفتاحية الموجودة في الجملة Si .
- $(KP_c \cap KP_{Si})$ هو عدد المفاهيم المفتاحية المشتركة بين الجملة Si والعنوان.

- $Similarity(Title, S_i)$ هي درجة تشابه العنوان مع جملة S_i التي تحسب بحساب تجيب الزاوية بين شعاع مفاهيم الجملة (بدلاً من شعاع مصطلحات الجملة) وشعاع مفاهيم العنوان باستخدام المعادلة (6-6).

$$Similarity_{Title, S_i} = \frac{\sum_{c=1}^n w_{cS_i} w_{cTitle}}{\sqrt{\sum_{c=1}^n w_{cS_i}^2} \sqrt{\sum_{c=1}^n w_{cTitle}^2}} \quad (6-6)$$

حيث:

- $Similarity_{Title, S_i}$: التشابه بين الجملة S_i والعنوان $Title$.
- n : عدد المفاهيم الفريدة في المستند .
- w_{cS_i} : وزن المفهوم c في الجملة S_i .
- w_{cTitle} : وزن المفهوم c في العنوان $Title$.

3-4-6 خاصية موقع الجملة:

تعتمد أهمية الجملة على موقعها في الفقرة / المستند بغض النظر عن مجال / موضوع المستند. اقترحت العديد من العلاقات فيما يتعلق بموقع الجملة في مجال تلخيص النصوص ((Gupta et al., 2011)، (Prasad et al., 2010)).

اعتمد في تحديد درجة خاصية موقع الجملة في نظامي التلخيص المقترحين في هذا البحث على فرضية أن الجملتين الأوليتين في المستند لها أهمية أكثر من باقي الجمل في المستند.

لذلك أعطيت $S_{Location}$ درجة تساوي 2، وأعطيت الجملتان الثالثة والرابعة درجة تساوي 1، وباقي الجمل أخذت درجة تساوي الصفر. وقد اختيرت هذه القيم تجريبياً للحصول على جمل موائمة.

4-4-6 خاصية طول الجملة

قد يؤثر طول الجملة $Sentence\ length$ على أهميتها. ومن ثم، قد تُستبعد الجمل الطويلة جداً أو القصيرة جداً من الملخص. اعتمدت إجرائية حساب خاصية طول الجملة في هذا البحث على طريقتين، طريقة تخص

التلخيص الإحصائي المقترح، وطريقة تخص التلخيص باعتماد الأنطولوجيا المقترح. نوضح فيما يلي طريقة حساب هذه الخاصية في النظامين.

1-4-4-6 خاصية طول الجملة في النظام الإحصائي:

تُحسب خاصية طول الجملة في التلخيص الإحصائي المقترح باستخدام المعادلة رقم (7-6)، وذلك بحساب حاصل قسمة عدد المصطلحات في الجملة على عدد المصطلحات في أطول جملة.

$$S_{length} = \frac{\text{number of words in the sentence}}{\text{number of words in the longest sentence}} \quad (7-6)$$

2-4-4-6 خاصية طول الجملة في التلخيص باعتماد الأنطولوجيا:

تُحسب خاصية طول الجملة في التلخيص باعتماد الأنطولوجيا بعد إعادة المصطلحات إلى مفاهيمها اعتماداً على الأنطولوجيا (وفقاً لخوارزمية إعادة مصطلحات الجمل المبينة في الشكل (2-6))، باستخدام المعادلة (8-6)، وذلك بحساب حاصل قسمة عدد المفاهيم في الجملة على عدد المفاهيم في أطول جملة.

$$S_{length} = \text{Length Score} = \frac{\text{number of concepts in the sentence}}{\text{number of concepts in the longest sentence}} \quad (8-6)$$

5-4-4-6 خاصية العبارات البارزة:

يمكن أن يكون وجود بعض العبارات البارزة مؤشراً جيداً على أهمية المحتوى ((Lakshmi et al., 2015)، ((Brazilay et al., 2015)). ومن ثم، تُعطى الجمل التي تحتوي على هذه العبارات درجة أعلى مقارنة بالجمل الأخرى.

تُحسب درجة هذه الخاصية في نظامي التلخيص المقترحين باستخدام المعادلة (9-6)، وذلك بحساب حاصل قسمة عدد العبارات البارزة في الجملة على عدد العبارات البارزة في المستند.

$$S_{Cue} = \frac{\text{number of Cue phrases in the sentence}}{\text{number of Cue phrases in the document}} \quad (9-6)$$

6-4-6 خاصية العبارات غير المهمة

بعض العبارات هي علامات للكلام بمثابة كلمات شرح. مثل هذه العبارات تضعف الجملة لأنها توجي بأن الجملة القادمة هي معلومات إضافية تتعلق بالفكرة الأساسية (Gupta et al., 2010). لذلك يجب تقليل درجة الجمل التي تحتوي على هذه العبارات.

تُحسب خاصية العبارات الضعيفة في نظامي التلخيص المقترحين باستخدام المعادلة رقم (10-6) وذلك بحساب مقلوب عدد الكلمات الضعيفة في الجملة.

$$S_{Weak} = \frac{1}{\text{number of words in the sentence}} \quad (10-6)$$

5-6 تقييم الجمل وتوليد الملخص:

يشكل الملخص النهائي في هذه المرحلة بتضمينه الجمل المهمة بترتيب ورودها في المستند الأساسي بنسبة التلخيص المطلوبة والمدخلة للنظام. وذلك وفقاً للمرحلتين هما: تثقيل الجمل، وترتيب الجمل وصياغة الملخص النهائي، وفيما يلي نقدم شرحاً لتلك المراحل.

تثقيل الجمل:

تُحسب درجة أهمية الجمل في التلخيص الإحصائي المقترح بناءً على المجموع الخطي لدرجات خصائص كل جملة في المستند استناداً إلى المصطلحات الواردة فيها (باستخدام الإجرائية sentence_weight المبينة في الملحق B) وفقاً للمعادلة (11-6) (Meena et al., 2016)، بينما تُستنتج درجة أهمية الجمل في المستند في التلخيص المعتمد على الأنطولوجيا المقترح استناداً إلى المجموع الخطي لخصائص كل جملة في المستند استناداً إلى مفاهيم المصطلحات الواردة فيها بدلاً من المصطلحات نفسها (وذلك باستخدام الإجرائية senetence_ontology_weight المبينة في الملحق B) وفقاً للمعادلة (11-6).

$$\text{Sentence Score} = S_{weak} * (S_{key} + S_{TitleSim} + S_{Location} + S_{length} + S_{cue}) \quad (11-6)$$

حيث:

- S_{key} : درجة خاصية العبارات المفتاحية المحسوبة بالمعادلة رقم (6-1) في التلخيص الإحصائي أوالمعادلة (6-2) في التلخيص باعتماد الأنطولوجيا،
- $S_{TitleSim}$: درجة خاصية التشابه مع العنوان المحسوبة بالمعادلة (6-3) في التلخيص الإحصائي أوالمعادلة (6-5) في التلخيص باعتماد الأنطولوجيا
- $S_{Location}$: درجة خاصية موقع الجملة المحسوبة حسب الفقرة خاصية موقع الجملة أعلاه.
- S_{length} : درجة خاصية طول الجملة المحسوبة بالمعادلة (6-7) في التلخيص الإحصائي أوالمعادلة رقم (6-8) في التلخيص باعتماد الأنطولوجيا.
- S_{cue} : درجة خاصية العبارات البارزة المحسوبة بالمعادلة (6-9).
- S_{weak} : درجة خاصية العبارات غير المهمة المحسوبة بالمعادلة (6-10).

ترتيب الجمل وصياغة الملخص:

تُرتب الجمل بعد حساب النتيجة الإجمالية ترتيباً تنازلياً بناءً على درجات ثقل الجمل المحسوبة بحساب المجموع الخطي لدرجات خصائصها وفقاً للمعادلة (6-11). ومن ثم تُختار الجمل ذات الترتيب الأعلى بناءً على نسبة التلخيص المطلوبة. ومن ثم يُعاد ترتيب الجمل المستخرجة المختارة بناءً على موضعها الأصلي في المستند للحفاظ على تناسق النص في الملخص الذي أنشئ. وبذلك يصاغ الملخص باختيار الجمل المهمة المرتبة حسب ورودها في المستند الأصلي (باستخدام الإجرائية `Ontology_summary` الموضحة في الملحق B).

الفصل السابع

تحليل النتائج وتقييمها

Results Analyses and Evaluation

7-1 مقدمة

عرض في الفصول السابقة كيفية بناء نظام التلخيص الآلي ، وكيفية تعلم الأنطولوجيا العربية المستخدمة في نظام التلخيص . يقدم هذا الفصل تحليلاً لنتائج التلخيص الآلي للنصوص العربية وتقييمها، كما يتناول شرحاً للأنطولوجيا العربية المتولدة.

7-2 تقييم التلخيص الآلي للنصوص العربية وتحليل نتائجه:

إن اختبار نظام تلخيص النصوص الآلي وتقييمه عملية صعبة نظراً لعدم وجود ملخص مثالي لمستند معين أو مجموعة من المستندات المرتبطة. علاوة على ذلك، فإن الافتقار إلى مجموعات البيانات القياسية العربية الحالية جعل عملية التقييم أكثر تعقيداً، وربما شخصية في بعض الحالات، لأن الباحثين يميلون إلى جمع مجموعات البيانات الخاصة بهم (Al-Saleh et al., 2016).

طبق التقييم باعتماد التوارد المشترك واستخدام مقاييس الدقة precision والاستنكار recall ومعامل F وهي المقاييس الأكثر شهرة واستخداماً (Oufaida et al., 2014).

تقيس الدقة مقدار صحة correction الملخص أو مقدار صحة المعلومات الموجودة في الملخص، وتحسب بتقاطع الجمل المستخرجة والجمل الموائمة مقسوماً على الجمل المستخرجة كلها باستخدام المعادلة (7-1).

$$P = \frac{\text{relevant Sentences} \cap \text{Retrieved Sentences}}{\text{retrieved Sentences}} \quad (1 - 7)$$

يُستخدم الاستدكار لقياس الاستكمال completeness أو مقدار تغطية المعلومات الموجودة في الملخص للمعلومات المهمة في المرجع المرتبط بالملخص. ويُحسب بنقاط الجمل المستخرجة والجمل المثالية مقسوماً على الجمل المثالية الموائمة كلها باستخدام المعادلة رقم (2-7).

$$R = \frac{\text{Relevant Sentences} \cap \text{Retrieved Sentences}}{\text{relavant Sentences}} \quad (2 - 7)$$

يُدمج المقياس F-Score بين الدقة Precision P والاستدكار Recall R (Salton, 1989) وهو متوسط مقاييس الدقة والاستدكار ويعطى بالمعادلة رقم (3-7).

$$F = \frac{2PR}{P+R} \quad (3-7)$$

7-2-1 مجموعة البيانات:

استُخدمت مجموعة الملخصات العربية من (Essex (EASC (El-Haj et al., 2010) لاختبار نظم التلخيص المقترحة وتقييمها. مجموعة EASC عبارة عن ملخص استخراجي من صنع الإنسان، نشر بواسطة مجموعة من الباحثين في جامعة إسيكس. يضم 153 مستنداً حول مواضيع مختلفة، جمعت من الصحف العربية وويكيبيديا. لكل مستند في مجموعة EASC هناك خمسة ملخصات مرجعية مختلفة أنشئت كل واحدة من قبل إنسان مختلف. تتميز مجموعة البيانات هذه بأنها مجموعة البيانات العربية الوحيدة المنشأة، والتي تجعل التقييم أكثر واقعية مقارنة بالمناهج الأخرى، مثل: الاعتماد على مجموعة بيانات مترجمة أو اعتماداً على خرج الملخصات المطورة مسبقاً.

ولحل مشكلة الشخونة في وجود خمس ملخصات مرجعية لكل مستند في مستندات EASC، أنشئ ملخص الأغلبية يسمى ملخص المعيار الذهبي الذي يضم الجمل الموجودة في معظم المراجع الخمسة (ثلاثة أو أكثر)، وهذا ما يسمى الملخص المرجعي الذهبي (El-Haj et al., 2011b). كما حددت عتبة نسبة التلخيص لإنشاء ملخص الخرج بـ 50%، وهي النسبة الوسطية لطول ملخص المعيار الذهبي المرجعي لكل مستند.

2-2-7 التجارب والنتائج:

حُسبت مقاييس الدقة والاستدكار والمعامل F وذلك باستخدام مجموعة الملخصات المرجعية EASC، من أجل تقييم نظام التلخيص الإحصائي المقترح، ونظام التلخيص باعتماد الأنطولوجيا المقترح، ونظام التلخيص المشهور لخصلي Lakhasly والمتوافر على الشبكة من خلال موقع ذي العنوان <https://lakhasly.com>، والذي يعد أحد أبرز المواقع التي تقوم بتلخيص الفقرات والنصوص العربية على الإنترنت Online وأكثرها استخداماً. يبين الشكل (1-7) عينة لأحد النصوص الملخصة وهو "لودفيج فان بيتهوفن" عدد جملة 54 جملة وعدد جمل الملخص المرجعي فيه 20 جملة. كما تبين الأشكال ((2-7)، ((3-7)، ((4-7) تلخيص هذا النص باستخدام نظام التلخيص باستخدام الأنطولوجيا والتلخيص الإحصائي وتلخيص نظام لخصلي على الترتيب، مع إظهار الجمل الموائمة بلون شفاف، ونسب الدقة والاستدكار ومعامل F في كل منها. كما يبين الجدول (3-9) (في الملحق A) جمل الملخص الناتجة عن التلخيص باعتماد الأنطولوجيا في مستند العينة مع قيم أوزانها ونسب خصائصها لهذا النص.

لودفيج فان بيتهوفن مؤلف موسيقي ألماني ولد عام 1770 م في مدينة بون. يعتبر من أبرز عباقرة الموسيقى في جميع العصور، وأبدع أعمالاً موسيقية خالدة. له الفضل الأعظم في تطوير الموسيقى الكلاسيكية. قدم أول عمل موسيقي وعمره 8 سنوات. تشمل مؤلفاته للأوركسترا تسعة سيمفونيات وخمس مقطوعات موسيقية على البيانو ومقطوعة على الكمان. كما ألف العديد من المقطوعات الموسيقية كمقدمات للأوبرا. بدأ بيتهوفن يفقد سمعه في الثلاثينيات من عمره إلا أن ذلك لم يؤثر على إنتاجه الذي ازداد في تلك الفترة وتميز بالإبداع. من أجمل أعماله السمفونية الخامسة والسادسة والتاسعة. وقد توفي في فيينا عام 1827م. شهدت مدينة بون الألمانية ميلاد الفنان العبقري لودفيج فان بيتهوفن في 16 ديسمبر عام 1770، وتم تعميده في 17 ديسمبر 1770. ظهر تميزه الموسيقي منذ صغره، فنشرت أولى أعماله وهو في الثانية عشر من عمره عام 1783 م. اتسعت شهرته كعازف بيانو في سن مبكرة، ثم زاد إنتاجه وذاع صيته كمؤلف موسيقي. عانى بيتهوفن كثيراً في حياته، عائلتيًا وصحياً، فبالرغم من أن أباه هو معلمه الأول الذي وجه اهتمامه للموسيقى ولقنه العزف على البيانو والكمان، إلا أنه لم يكن الأب المثالي، فقد كان مدمناً للكحول، كما أن والدته توفيت وهو في السابعة عشر من عمره بعد صراع طويل مع المرض، تاركة له مسؤولية العائلة. مما منعه من إتمام خطته والسفر إلى فيينا، عاصمة الموسيقى في ذلك العصر. فهل كان التأليف الموسيقي هو نوع من أنواع العلاج والتغلب على المشاكل بالنسبة لبيتهوفن. في 1789 م تحقق حلمه أخيراً، فقد أرسله حاكم بون إلى فيينا، وهناك تتلمذ على يد هايدن. ولكن بيتهوفن، صاحب الألحان واجه بعض الخلافات مع معلمه، وعندما سافر هايدن إلى لندن، تحول بيتهوفن إلى معلمين آخرين مثل ساليري وشينك وألبريشنبرجر. وقد أسهمت كل هذه الدروس والاحتكاكات في تكوين شخصية بيتهوفن الفنية. وحاول أن يشق لنفسه طريق كعازف في عاصمة الموسيقى، وسرعان ما لاقى مكانة كبرى خاصة في الأوساط الأرستقراطية. فقد حاز بإعجاب الأسرة الملكية وعومل كصديق أكثر منه مؤلفاً. بالرغم من ذلك فقد عاش ومات فقيراً، غناه هو أعماله الفنية المتميزة. فقد جاء إنتاجه الفني غزيراً حتى بعد إصابته بالصمم. بدأت إصابة بيتهوفن بصمم بسيط عام 1802، فبدأ في الانسحاب من الأوساط الفنية تدريجياً، وأمضى حياته بلا زواج يرتبط بعلاقات عدة مع سيدات صغيرات. إلا أنه لم يتوقف عن الإنتاج الفني، ولكن أعماله اتخذت اتجاه جديد. ومع ازدياد حالة الصمم التي أصابته، امتنع عن العزف في الحفلات العامة، وابتعد عن الحياة الاجتماعية واتجه للوحدة، وقلت مؤلفاته، وأصبحت أكثر تعقيداً. حتى أنه رد على انتقادات نقاده بأنه يعزف للأجيال القادمة. وبالفعل مازالت أعماله حتى اليوم من أهم ما أنتجته الموسيقى الكلاسيكية العالمية. واكتسبت اثنان من السيمفونيات التي كتبها في صممه أكبر شعبية، وهما السيمفونية الخامسة والتاسعة. كما أنه أحدث الكثير من التغييرات في الموسيقى، وأدخل الغناء والكلمات في سيمفونيته التاسعة. فجاءت رسالته إلى العالم كل البشر سيصبحون إخوة.

الشكل (1-7) عينة النص الأصلي (54 جملة ، 20 جملة موائمة ملونة بلون شفاف)

لودفيج فان بيتهوفن مؤلف موسيقي ألماني ولد عام 1770 م في مدينة بون، يعتبر من أبرز عباقرة الموسيقى في جميع العصور، وأبدع أعمالاً موسيقية خالدة، له الفضل الأعظم في تطوير الموسيقى الكلاسيكية، تشمل مؤلفاته للأوركسترا تسعة سيمفونيات وخمس مقطوعات موسيقية على البيانو ومقطوعة على الكمان، بدأ بيتهوفن يفقد سمعه في الثلاثينيات من عمره إلا أن ذلك لم يؤثر على إنتاجه الذي ازداد في تلك الفترة وتميز بالإبداع، وقد توفي في فيينا عام 1827م، شهدت مدينة بون الألمانية ميلاد الفنان العبقري لودفيج فان بيتهوفن في 16 ديسمبر عام 1770، ظهر تميزه الموسيقي منذ صغره، فنشرت أولى أعماله وهو في الثانية عشر من عمره عام 1783 م، ثم زاد إنتاجه وذاع صيته كمؤلف موسيقي، فبالرغم من أن أباه هو معلمه الأول الذي وجه اهتمامه للموسيقى ولقنه العزف على البيانو والكمان، عاصمة الموسيقى في ذلك العصر، فهل كان التأليف الموسيقي هو نوع من أنواع العلاج والتغلب على المشاكل بالنسبة لبيتهوفن، فقد أرسله حاكم بون إلى فيينا، وهناك تتلمذ على يد هايدن ولكن بيتهوفن، تحول بيتهوفن إلى معلمين آخرين مثل ساليري وشينك وألبريشتبيرجر، وحاول أن يشق لنفسه طريق كعازف في عاصمة الموسيقى، غناه هو أعماله الفنية المتميزة، بدأت إصابة بيتهوفن بصمم بسيط عام 1802، وأمضى حياته بلا زواج يرتبط بعلاقات عدة مع سيدات صغيرات، ومع ازدياد حالة الصمم التي أصابته، امتنع عن العزف في الحفلات العامة، وابتعد عن الحياة الاجتماعية واتجه للوحدة وقلت مؤلفاته، وأصبحت أكثر تعقيداً، وبالفعل مازالت أعماله حتى اليوم من أهم ما أنتجته الموسيقى الكلاسيكية العالمية، كما أنه أحدث الكثير من التغييرات في الموسيقى

الشكل (7-2) الملخص باستخدام الأنطولوجيا (الدقة 0.57 ، الاستدكار 0.75 ومعامل 0.65 F)

لودفيج فان بيتهوفن مؤلف موسيقي ألماني ولد عام 1770 م في مدينة بون، يعتبر من أبرز عباقرة الموسيقى في جميع العصور، وأبدع أعمالاً موسيقية خالدة، له الفضل الأعظم في تطوير الموسيقى الكلاسيكية، قدم أول عمل موسيقي وعمره 8 سنوات، تشمل مؤلفاته للأوركسترا تسعة سيمفونيات وخمس مقطوعات موسيقية على البيانو ومقطوعة على الكمان، كما ألف العديد من المقطوعات الموسيقية كمقدمات للأوبرا، بدأ بيتهوفن يفقد سمعه في الثلاثينيات من عمره إلا أن ذلك لم يؤثر على إنتاجه الذي ازداد في تلك الفترة وتميز بالإبداع، شهدت مدينة بون الألمانية ميلاد الفنان العبقري لودفيج فان بيتهوفن في 16 ديسمبر عام 1770، وتم تعميده في 17 ديسمبر 1770، فنشرت أولى أعماله وهو في الثانية عشر من عمره عام 1783 م، ثم زاد إنتاجه وذاع صيته كمؤلف موسيقي، عانى بيتهوفن كثيراً في حياته عائلياً وصحياً، فبالرغم من أن أباه هو معلمه الأول الذي وجه اهتمامه للموسيقى ولقنه العزف على البيانو والكمان، عاصمة الموسيقى في ذلك العصر، فهل كان التأليف الموسيقي هو نوع من أنواع العلاج والتغلب على المشاكل بالنسبة لبيتهوفن، فقد أرسله حاكم بون إلى فيينا، وهناك تتلمذ على يد هايدن ولكن بيتهوفن، وقد أسهمت كل هذه الدروس والاحتكاكات في تكوين شخصية بيتهوفن الفنية، وحاول أن يشق لنفسه طريق كعازف في عاصمة الموسيقى، بدأت إصابة بيتهوفن بصمم بسيط عام 1802، وأمضى حياته بلا زواج يرتبط بعلاقات عدة مع سيدات صغيرات، إلا أنه لم يتوقف عن الإنتاج الفني، امتنع عن العزف في الحفلات العامة، وبالفعل مازالت أعماله حتى اليوم من أهم ما أنتجته الموسيقى الكلاسيكية العالمية، وهما السيمفونية الخامسة والتاسعة، كما أنه أحدث الكثير من التغييرات في الموسيقى

الشكل (7-3) الملخص باستخدام التلخيص الإحصائي (الدقة 0.45 والاستدكار 0.6 ومعامل 0.51 F)

لودفيج فان بيتهوفن مؤلف موسيقي ألماني ولد عام 1770 م في مدينة بون. وأبدع أعمالاً موسيقية خالدة. تشمل مؤلفاته للأوركسترا تسعة سيمفونيات وخمس مقطوعات موسيقية على البيانو ومقطوعة على الكمان. كما ألف العديد من المقطوعات الموسيقية كمقدمات للأوبرا. بدأ بيتهوفن يفقد سمعه في الثلاثينيات من عمره إلا أن ذلك لم يؤثر على إنتاجه الذي ازداد في تلك الفترة وتميز بالإبداع. شهدت مدينة بون الألمانية ميلاد الفنان العبقري لودفيج فان بيتهوفن في 16 ديسمبر عام 1770، فنشرت أولى أعماله وهو في الثانية عشر من عمره عام 1783 م. ثم زاد إنتاجه وذاع صيته كمؤلف موسيقي. عانى بيتهوفن كثيراً في حياته، فبالرغم من أن أباه هو معلمه الأول الذي وجه اهتمامه للموسيقى ولقنه العزف على البيانو والكمان، كما أن والدته توفيت وهو في السابعة عشر من عمره بعد صراع طويل مع المرض، فهل كان التأليف الموسيقي هو نوع من أنواع العلاج والتغلب على المشاكل بالنسبة لبيتهوفن. فقد أرسله حاكم بون إلى فيينا، وهناك تتلمذ على يد هايدن. وعندما سافر هايدن إلى لندن، تحول بيتهوفن إلى معلمين آخرين مثل ساليري وشينك وألبريشتبيرجر. وقد أسهمت كل هذه الدروس والاحتكاكات في تكوين شخصية بيتهوفن الفنية. وحاول أن يشق لنفسه طريق كعازف في عاصمة الموسيقى، غناه هو أعماله الفنية المتميزة. فقد جاء إنتاجه الفني غزيراً حتى بعد إصابته بالصمم. بدأت إصابة بيتهوفن بصمم بسيط عام 1802، امتنع عن العزف في الحفلات العامة، وابتعد عن الحياة الاجتماعية واتجه للوحدة، وبالفعل مازالت أعماله حتى اليوم من أهم ما أنتجته الموسيقى الكلاسيكية العالمية. فجاءت رسالته إلى العالم كل البشر سيصبحون إخوة

الشكل (7-4) الملخص باستخدام لخصلي Lakhasly (الدقة 0.4 والاستدكار 0.5 ومعامل 0.44 F)

7-2-3 تحليل النتائج وتقييمها:

بعد دراسة وتحليل نتائج التقييم على نصوص الملخصات المرجعية تبين أن متوسط الدقة والاستدكار ومعامل F في التلخيص باعتماد الأنطولوجيا المقترح هي: 0.74 و 0.85 و 0.78 على الترتيب. وكانت في التلخيص الإحصائي المقترح هي: 0.64 و 0.73 و 0.67 على الترتيب. وفي نظام تلخيص لخصلي هي: 0.58 و 0.59 و 0.55 على الترتيب. وهي تتفاوت حسب طول المستند.

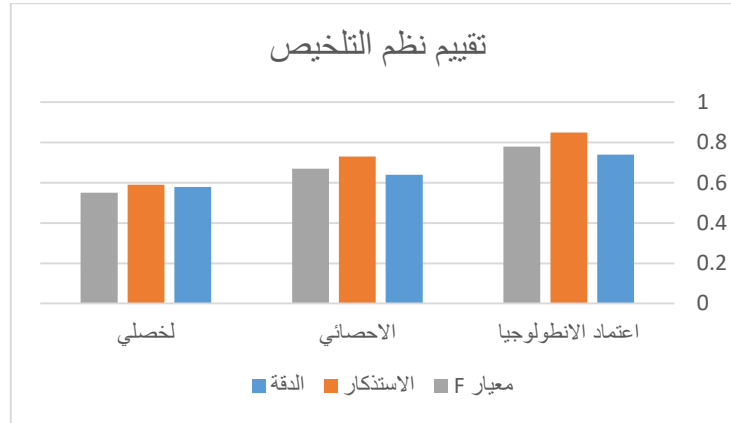
وقد أوضحت النتائج كما هو مبين في الجدول (7-1) أن الدقة والاستدكار ومعامل F في التلخيص المعتمد على الأنطولوجيا أفضل من النظام الإحصائي المقارن معه وتتفوق عليه بنسب 0.1 و 0.12 و 0.11 على الترتيب، وتفوق أيضاً على نتائج نظام لخصلي بنسب 0.16 و 0.26 و 0.23 على الترتيب. وتبين أيضاً أن النظامين المقترحين في هذا المشروع أفضل من نظام لخصلي، إلا أن نظام التلخيص المعتمد على الأنطولوجيا هو الأفضل. وهذا موضح في المخطط البياني في الشكل (7-5). باختصار، نظام التلخيص باعتماد الأنطولوجيا المقترح متفوق من حيث الدقة والاستدكار ومعامل F على طرائق التلخيص الأخرى المقارن معها. كما أن نتائج التقييم هذه جيدة مقارنة مع نظم التلخيص في مشاريع تلخيص النصوص العربية عموماً.

7-3 تقييم الأنطولوجيا العربية المتولدة وتحليل نتائجها:

وُلدت الأنطولوجيا العربية بتوسيع شبكة الكلمات العربية من خلال خصائص الموسوعة العربية باستخدام تقنية التقابل، اعتمد في تقييمها على التقييم باعتماد التطبيق، وذلك من خلال اعتماد الأنطولوجيا في نظام التلخيص المقترح في هذا البحث، ومقارنة نتائج تلخيص نظام التلخيص باعتماد الأنطولوجيا مع نتائج تلخيص نظام التلخيص غير المعتمد على الأنطولوجيا، فبينت النتائج حسب الجدول (7-1) أن التلخيص باستخدام الأنطولوجيا أفضل من حيث معايير الدقة والاستدكار. وفيما يلي نبين مجموعة البيانات المستخدمة والنتائج التي توصل إليها.

الجدول (1-7) مقارنة مقاييس الدقة والاستدكار والمعامل F لأنظمة التلخيص الثلاثة (التلخيص باعتماد الأنطولوجيا، والتلخيص الإحصائي، وتلخيص نظام لخصلي)

الدقة	الاستدكار	معيار F
0.74	0.85	0.78
0.64	0.73	0.67
0.58	0.59	0.55



الشكل (5-7) مقارنة نتائج تقييم التلخيص باعتماد الأنطولوجيا مع التلخيص الإحصائي وتلخيص لخصلي.

7-3-1 مجموعة البيانات

استخدمت الموسوعة العربية بوصفها مصدراً معرفياً غنياً يضم 15000 مقالة مكتوبة بلغة عربية صحيحة في مجالات المعرفة الإنسانية كافة. فقد فُهرست نصوصها فهرسة آلية لتوليد خصائص الموسوعة العربية، واستخرجت علاقات التشابه والترافق بطريقة إحصائية اعتماداً على توارد المصطلحات في نصوص الموسوعة وتكرارها (كما ذكر في الفصل الخامس)

استخدمت شبكة الكلمات العربية والتي تتألف من 11270 مفهوماً و 23496 مصطلحاً بعد فهرستها فهرسة آلية من أجل تجهيزها لعملية التقابل بينها وبين خصائص الموسوعة العربية (كما ذكر في الفصل الخامس)

7-3-2 نتائج الأنطولوجيا العربية المتولدة:

طبقت خوارزمية التعلم لتوسيع شبكة الكلمات العربية بإغنائها بمفاهيم ومصطلحات وعلاقات مستخرجة من

الموسوعة العربية وتوصل البحث إلى:

- عدد المفاهيم المضافة إلى شبكة الكلمات 5778 مفهوماً.
- عدد المصطلحات المضافة 10212 مصطلحاً.
- عدد العلاقات المضافة بين المصطلحات من نوع ترافق: 3500 علاقة.
- عدد العلاقات المضافة بين المصطلحات من نوع تشابه: 2400 علاقة.

يبين الجدول (9-4) (في الملحق A) عينات من مجموعات المفاهيم المجردة والمصطلحات المجردة المرتبطة

في الأنطولوجيا العربية. كما يبين الجدول (9-5) (في الملحق A) عينات من علاقات التشابه والترافق بين المصطلحات المجردة في الأنطولوجيا العربية.

7-4 تطبيق التلخيص الآلي على الموسوعة العربية:

طبّق التلخيص الآلي باعتماد الأنطولوجيا على مقالات الموسوعة العربية بنسبة تلخيص 60% ، وحصل

على ملخصات جيدة لتلك المقالات متضمنة الجمل الأكثر أهمية وفقاً لمعايير المستخدمة في الملخص الآلي. وقد

اخترنا عدداً من مقالات الموسوعة العربية، من اختصاصات متنوعة، وبأطوال متفاوتة بين الصفحة إلى أربع

صفحات.

الفصل الثامن

الخلاصة والخاتمة

Summary and Conclusion

8-1 مقدمة:

عرض في الفصول السابقة كيفية بناء الملخص الآلي للنصوص العربية اعتماداً على أنطولوجيا عربية وُلدت بتوسيع شبكة الكلمات العربية من خلال خصائص الموسوعة العربية. كما عرضت النتائج وتقييم النظام. نقدم في هذا الفصل خلاصة ما أنجز في هذا البحث، والخاتمة، والآفاق المستقبلية التي يمكن بها التوسع في المشروع المقدم في هذا البحث وتطويره مستقبلاً.

8-2 الخلاصة:

أصبح تلخيص المستندات من القضايا المهمة في ظل هذا الكم الهائل من المحتوى الرقمي الموجود الآن على الإنترنت وبخاصة المحتوى النصي، حيث يميل المستخدم عادةً إلى استخراج أهم المعلومات أو معرفة ما يشير إليه المصدر الأساسي للمعلومات. يستطيع التلخيص الآلي للنصوص توفير الوقت والجهد للمستخدم، فيستطيع الأخير اتخاذ القرارات بشكل أسرع. لذا أصبح هناك اهتمام بالتلخيص الآلي للنصوص بشكل عام وللنصوص العربية بشكل خاص لأنها اللغة الأم لأكثر من 300 مليون شخص حول العالم وهي واحدة من اللغات الست الرسمية لدى الأمم المتحدة (Farghaly et al., 2009). ولكن الأبحاث والدراسات في مجال تلخيص النصوص العربية ضئيلة، وتتمو ببطء (El-Haj et al., 2011a) مقارنة مع اللغات العالمية الأخرى مثل اللغة الإنكليزية، كما اتبعت في معظمها طرائق استخراجية، وابتعدت عن الأسلوب التجريدي، ولم تستخدم مصادر معرفية مثل الأنطولوجيا لتحسين جودة التلخيص الآلي من خلال العلاقات الدلالية بين المفاهيم.

لذا اقترح في هذا البحث نموذج لإيجاد ملخص آلي لتلخيص النصوص العربية باستخدام الأنطولوجيا، ونظام تلخيص مبني على النموذج المقترح الذي تضمن: توليد أنطولوجيا عربية بالتقابل بين شبكة الكلمات العربية Arabic Wordnet وخصائص مقالات الموسوعة العربية، وتوليد ملخص آلي يعتمد على الأنطولوجيا، واختبار التلخيص الدلالي المقترح وتحليل النتائج وتقييمها ومن ثم تطبيق هذا النموذج على مقالات الموسوعة العربية لتلخيصها.

وُلِدَت أنطولوجيا عربية من أجل الاعتماد عليها في نظام التلخيص الآلي، وذلك بتوسيع شبكة الكلمات العربية المترجمة، بإغنائها بمفاهيم ومصطلحات عربية، وعلاقات دلالية مستنتجة من خصائص الموسوعة العربية، باستخدام تقنية التقابل، اعتماداً على كون الموسوعة العربية من أغنى المصادر المعرفية العربية.

بُنِيَ نظام تلخيص استخراجي عام أحادي اللغة وأحادي المستند يستخدم تقنيات دلالية بالاعتماد على الأنطولوجيا، موظفاً خصائص إحصائية لتثقيل الجمل وتقييم أهميتها استناداً إلى توارد مفاهيم المصطلحات الواردة فيها وتكرارها بدلاً من المصطلحات نفسها باستخدام علاقات دلالية بين المصطلحات تستنبط من الأنطولوجيا العربية المقترحة، وهي: الترادف، والاشتغال، والتضمين، والعضو. وبذلك يتم الدمج بين الطريقة الإحصائية والدلالية في التلخيص الآلي من أجل تحسين جودة التلخيص.

ولاختبار نظام التلخيص استخدمت مجموعة الملخصات العربية من (El-Haj et al.,) Essex (EASC) 2010)، وطُبق التقييم باعتماد التوارد المشترك واستخدام مقاييس الدقة precision والاستدكار recall ومعامل F وهي المقاييس الأكثر شهرة واستخداماً (Oufaida et al., 2014). فبينت النتائج أن متوسط الدقة والاستدكار ومعامل F في التلخيص باعتماد الأنطولوجيا المقترح هي: 0.74 و0.85 و0.78 على التوالي. وهي نسب جيدة مقارنة مع نسب أنظمة التلخيص الآلي.

3-8 الخاتمة:

طُور في هذا البحث نظام تلخيص عام استخراجي للنصوص العربية باعتماد الأنطولوجيا، مُثِلت فيه جمل النصوص المراد تلخيصها بمفاهيمها اعتماداً على العلاقات الدلالية في الأنطولوجيا العربية، وقُيِّمت أهميتها اعتماداً على حساب المجموع الخطي لدرجات خصائص مفاهيمها، وبذلك دمجت تقنيات التلخيص الإحصائية والدلالية لتحسين جودة التلخيص. فقد بينت نتائج التقييم أن الملخصات الناتجة في هذا النظام أفضل من نتائج ملخصات الأنظمة المقارن معها. لذا يمكن أن يسهم هذا النظام في خدمة القارئ العربي واللغة العربية بتوفير ملخصات للنصوص العربية ذات جودة جيدة.

كما وُلدت هذا البحث أنطولوجيا عربية لاعتمادها في التلخيص الآلي، التي تعد بحد ذاتها عملاً مهماً ويسهم في دعم التطبيقات التي تعتمد معالجة اللغة العربية، مثل: الوب الدلالي، والترجمة الآلية، والتلخيص الآلي وغيرها.

4-8 الآفاق المستقبلية:

يمكن أن تبحث الدراسات المستقبلية باستكمال نظام التلخيص باعتماد الأنطولوجيا المقترح في هذا البحث بتحسين نتائجه بإلغاء التكرار، باستخدام طرائق تجميع الجمل المتشابهة، وإضافة خصائص أخرى تسهم في تحسين أوزان الجمل، التي تعكس أهمية تضمينها في الملخص النهائي. إضافة إلى ذلك يمكن أن يطور النظام لتلخيص مستندات متعددة، كما يمكن أن يُطوّر نظام التلخيص مستقبلاً بالتحول إلى اعتماد التلخيص التجريدي بدلاً من الاستخراجي بالاستفادة من التمثيل الدلالي لجمل المستند بالاستعانة بالأنطولوجيا العربية.

علاوة على ذلك، يمكن استكمال مشروع التلخيص الآلي للنصوص العربية باستخدام الأنطولوجيا ليصبح أداة تلخيص دلالي تستخدم في موقع خاص على الإنترنت يعتمد لتلخيص آلي متعدد المستندات وبلغات متعددة.

ومن جانب آخر، تعد الأنطولوجيا العربية المتولدة نواة لأنطولوجيا عربية شاملة يمكن تطويرها واستكمالها بإضافة مفاهيم وعلاقات جديدة إليها باستخدام تقنية التقابل مع خصائص مصادر معرفية عربية موثقة أخرى مثل المعاجم العربية، للحصول على أنطولوجيا عربية أكثر فاعلية وفعالية.

المراجع والمصادر

REFERENCES

- Abdelkrime A., Djamel Eddine Z., Khaled Walid H., "**Allsummarizer system at multiling multilingual single and multi-document summarization.**" SIGDIAL 2015 Conference. 2015. 237–244.
- Abuobieda A., Salim N., Albaham A., Osman A., Kumar Y., "**Text summarization features selection method using pseudo genetic-based model.**" International conference on information retrieval knowledge management. 2012. 193–197.
- Ajith A., Swagatam D., Sandip R., "**Swarm Intelligence Algorithms for Data clustering.**" Soft Computing for Knowledge Discovery and Data Mining, , Springer US,, 2008. 279-313.
- Al khalifa M., Rodríguez H., "**Automatically Extending Named Entities coverage of Arabic WordNet using Wikipedia.**" International Journal on Information and Communication Technologies, 2008.
- Al Maleh M., Desouki S., "**Arabic text summarization using deep learning approach.**" Journal of Big Data 7: 109, 2020.
- Al-Abdallah R.Z., Al-Taani A.T., "**Arabic single-document text summarization using particle swarm optimization algorithm.**" Procedia Computer Science 117, 2017: 30–37.
- Alami N., Meknassi M., Ouatik S.A., Ennahnahi N., "**Impact of stemming on arabic text summarization.**" 4th IEEE International Colloquium on Information Science and Technology (CiSt). IEEE. 2016.
- AlBarhamtoshy H., Wajdi A., "Designing and Implementing Arabic WordNet Semantic-Based." Language Engineering Conference. 2009. 23 -32.
- Aliane H., Alimazighi Z. Ahmed Cherif M., "**Al-Khalil: The Arabic Linguistic Ontology Project**", . Research Center on Scientific and technical Information,, Algiers.: Semantic web and Arabic Language Team, in Computer Science Department, University, 2009.
- AL-Khawaldeh F., Samawi V., "**Lexical cohesion and entailment based segmentation for arabic text summarization (Iceas).**" World Comput Science Information Technology Journal 5(3):51–60, 2015.
- AlMaayah M., Sawalha M., Abushariah M., "**Towards an automatic extraction of synonyms for Quranic Arabic WordNet.**" International Journal of Speech Technology, Volume 19, 2016: 177–189.
- Alotaiby F, Foda S. Alkharashi I., "**New approaches to automatic headline generation for arabic documents.**" Journal of Engineering Computer Innovation 3(1):11–25, 2012.
- Al-Saleh A., Menail M., "**Automatic arabic text summarization: a survey. Artif .**" Intell. Rev. Arch. 45 (2), 2016: 203–234.
- Al-Salhi R.Y., Abdullah A.M., "**Building Quranic stories ontology using MappingMaster domain-specific language.**" International Journal of Electrical and Computer Engineering (IJECE), 2022.
- Alsudias L., Rayson P., "**Developing an Arabic Infectious Disease Ontology to Include Non-Standard Terminology .**" Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020). 2020. 4842–4850.
- Althobaiti M., Kruschwitz U., Poesio M., "**Aranlp: A java-based library for the processing of arabic text.**" 2014.
- AlZahrani A., Mathkour H., Ismail Abdalla H., "**Pso-based feature selection for Arabic text summarization.**" Journal UCS, 21(11), 2015: 1454–1469.
- Anjali D. R., Lobo L., "**Text Summarization using Clustering Technique.**" International Journal of Engineering Trends and Technology (IJETT), 2013: 3348-3351.
- Aone C, Okurowski M. E. Gorfinsky J., "**Trainable, scalable summarization using robust NLP and machine learning.**" 17th International Conference on Computational Linguistics. Stroudsburg PA, USA, 1998. 62–66.

- Asim M., Wasim M. Khan M.U.G., "**A survey of ontology learning techniques and applications.**" Database, 2018.
- Ayedh A., Tan G., Alwesabi K., Rajeh H., "**The effect of preprocessing on arabic document categorization**" Algorithms 9 (2), 2016.
- Azmi AM, Al-Thanyyan S., "**A text summarizer for arabic.**" Computer Speech Language 26(4):260–273, 2012.
- Bawakid A., "**Automatic documents summarization using ontology based methodologies.**" Doctoral Thesis., University of Birmingham, 2011.
- Baxendale P., "**Man-made index for technical literature - an experiment.**" IBM Journal of Research and Development, 1958: vol. 2, no. 4.
- Bazghandi M., Tabrizi Gh T. Vafaei Jahan M., "**Extractive Summarization of Farsi Documents Based On PSO Clustering.**" International Journal of Computer Science Issues, 2012: Vol. 9, Issue 4, No 3.
- Belguith L, Ellouze M. Maaloul M. Jaoua M. Jaoua F. Blache P., "**Automatic summarization. In: Zitouni I (ed) Natural language processing of semitic languages, theory and applications of natural language processing.**" Springer, Berlin, 2014: 371–408.
- Belkebir R., Guessoum A., "**A supervised approach to arabic text summarization using adaboost.**In: Rocha A, Correia AM, Costanzo S, Reis LP (eds) **New contributions in information systems and technologies.**" advances in intelligent systems and computing, 2015: vol 353, 227–236.
- Belkredim Z., Meziane F., "**DEAR-ONTO: A DERivational ARabic Ontology Based on Verbs.**" International Journal of Computer Processing of Languages, 21 (3), 2008: 279-291.
- Benassi, Bergamaschi S., Vincini M., "**Web Semantic Search with TUCUXI.**" the Twelfth Italian Symposium on Advanced Database Systems SEBD., 2004. 426–433.
- Biemann C. B., "**Ontology learning from text: A survey of methods.**" LDV Forum 20(2), 2005: 75–93.
- Binwahlan M. S, Salim S. Suanmali S., "**Swarm Based Features Selection for Text Summarization.**" IJCSNS International Journal of Computer Science and Network Security, vol. 9, no. 1, , 2009: 175-179.
- Black W., Farwell D. Elkateb., "**Arabic WordNet and the Challenges of Arabic.**" Arabic NLP/MT Conference. London, U.K., 2006.
- Boudabous M, Maaloul M. Belguith L., "**Digital learning for summarizing arabic documents. In: Loftsson H, Rgnvaldsson E, Helgadtir S (eds) Advances in natural language processing, lecture notes in computer science.**" Springer, Berlin, 2010: 79-84.
- Brandow R., Mitze K. Rau L F., "**Automatic condensation of electronic publications by sentence selection.**" Information Processing & Management, 1995: 675-685.
- Brazilay R., Elhadad, M., "**Using lexical chains for text summarization.**" Text Summarization. 2015. 111–121.
- Brown J. D. B., "**Developing an automatic document classification system: A review of current literature and future directions.**" Technical Report, Canada: Defence R&D Canada - Ottawa Ottawa ONT (CAN), 2010.
- Burges C., Shaked T. Renshaw E. Deeds M. Hamilton N. Hullender G., "**Learning to rank using gradient descent.**" ICML. 2005. 89--96.
- Carenini G., Cheung J. C. K., "**Extractive vs. NLG-based abstractive summarization of evaluative text: the effect of corpus controversy.**" Fifth International Natural Language Generation Conference. PA, USA: Stroudsburg, 2008. 33–41.
- Chandra M., Gupta V., Paul S., "**A Statistical approach for Automatic Text Summarization by Extraction.**" International Conference on Communication Systems and Network Technologies. 2011. 268-271.
- Chen Y L., Hung L., "**Using decision trees to summarize associative classification rules.**" Expert Systems with Applications: An International Journal, 2009: 2338–2351.
- Chengcheng Li., "**Automatic Text Summarization based on Rhetorical Structure Theory.**" ICCASM (ICCASM.2010.5622918), 2010.
- Conroy J. M., O'leary D. P., "**Text summarization via hidden Markov models.**" 24th annual international ACM SIGIR conference on Research and development in information retrieval. New York, 2001. 406–407.

- Conroy J., Judith S., **"CLASSY Query-Based Multi-Document Summarization."** Document Understanding Conference (DUC-2005). Vancouver, Canada, 2005.
- Cristea D, Postolache O. Puscasu G. Ghetu L., **"Summarizing documents based on cue-phrases and references."** International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization. Italy, 2003.
- Daume H., Marcu D., **"Bayesian Multi-Document Summarization at MSE."** Workshop on Multilingual Summarization Evaluation (MSE). 2005.
- Deerwester S., Dumais S. Furnas G. Landauer T. Harshman R., **"Indexing by latent semantic analysis."** Journal Of The American Society For Information Science., 1990: 391-407.
- Dorr B., Zajic D. Schwartz R., **"Hedge Trimmer: A Parse-and-Trim Approach to Headline Generation."** HLT-NAACL 2003 Text Summarization Workshop. 2003.
- Edmundson H. P., **"New Methods in Automatic Extracting."** ACM, 1969: 264–285.
- El-Beltagy S., Rafea A., **"Kp-miner: a keyphrase extraction system for english and arabic documents. ."** Inf Syst 34 (1), 132–144, 2009.
- El-Fishawy N, Hamouda A. Attiya GM. Atef M., **"Arabic summarization in twitter social network."** Ain Shams Eng J 5(2):411–420, 2013.
- El-Ghannam F, El-Shishtawy T., **"Multi-topic multi-document summarizer."** International Journal of Computer Science Inf Technol 5(6):77–90, 2013.
- Elghazaly T., Ibrahim A., **"Arabic text summarization using rhetoricalstructure theory."** International Conference on INFormatics and Systems. 2012.
- El-Haj M, Kruschwitz U Fox C., **"multilingual summarisation pilot."** the text analysis conference (TAC) 2011. Maryland: University of essex at the tac 2011, 2011c.
- El-Haj M., Kruschwitz U, Fox C., **"Exploring clustering for multi-document arabic summarisation."** Salem M, Shaalan K, Oroumchian F, Shakery A, Khelalfa H (eds) Information retrieval technology, lecture notes in computer science, vol 7097. Springer, Berlin, 2011a: 550-561.
- El-Haj M., Kruschwitz U. Fox C., **"Multi-document arabic text summarisation."** Computer science and electronic engineering conference (CEEC), 2011 3rd, IEEE Xplore, 2011b: 40–44.
- El-Haj M., Kruschwitz Udo, Fox Ch. **"Using mechanical turk to create a corpus of arabic summaries."** 2010.
- El-Haj M., Rayson P., **"Using a keyness metric for single and multi document summarisation."** Proceedings of the MultiLing 2013 workshop on multilingual multi-document summarization. Sofia, Bulgaria: Association for Computational Linguistics, 2013. 64-71.
- El-Shishtawy T., El-Ghannam F., **"keyphrase based Arabic summarizer (KPAS)."** 8th International Conference, IEEE. informatics and systems (INFOS), 2012.
- Erkan G., Radev D., **"Lexrank: Graph-based lexical centrality as salience in text summarization."** J. Artificial Intelligence Research 22, 2004: 457–479.
- Fachrurrozi M., Utami R. Novi Yusliani Y., **"Frequent Term Based Text Summarization for Bahasa Indonesia."** Conference: International Conference on Innovations in Engineering and Technology (ICIET'2013). Bangkok, Thailand, 2013.
- Fan W., Wallace L. Zhongj S. R., **"Tapping into the Power of Text Mining."** ACM, 2005: 76-82.
- Farghaly A., Shaalan K. , **"Arabic Natural Language Processing: Challenges and Solutions"**, ACM Transactions on Asian Language Information Processing, Vol. 8, No. 4, Article 14, Pub. date: December 2009.
- Fattah MA, Ren F., **"Ga, mr, ffnn, pnn and gmm based models for automatic text summarization."** Computer Speech Language 23(1):126–144, 2009.
- Fawcett T. E., **"Feature discovery for problem solving systems"**. PhD Thesis, Amherst, MA, USA: University of Massachusetts, 1993.
- Fejer H., Omar N., **"Automatic arabic text summarization using clustering and keyphrase extraction."** 2014 International Conference on information technology and multimedia (ICIMU). 2014. 293–298.

- Fellbaum C., **"WordNet: an electronic lexical database."** Cambridge Mass: MIT Press, 1998.
- Fernando S., Stevenson M., **"Mapping WordNet synsets to Wikipedia articles"**. University of Sheffield, 2009.
- Ferreira R., Cabral L., Dueire R., Silva G., Freitas F., Cavalcanti G., **"Assessing sentence scoring techniques for extractive text summarization."** Expert System Application 40, 2012: 5755–5764.
- Froud H, Lachkar A. Ouati kSA., **"Arabic text summarization based on latent semantic analysis to enhance arabic documents clustering."** International Journal of Data Mining Knowledge Management Processing (IJDKP) 3(1). doi:10.5121/ijdkp. 2013.3107, 2013.
- Gomaa W., Fahmy A., **"A survey of text similarity approaches. ."** International Journal of Computer Application 68 (13), 2013.
- Gong Y., Liu X., **"Generic Text Summarization using Relevance Measure and Latent Semantic Analysis."** the Annual. International ACM SIGIR Conference on Research and Development in Information Retrieval., Louisiana, USA: New Orleans, 2011. 19-25.
- Gupta P., Pendluri V., Vats I., **"Summarizing text by ranking text units according to shallow linguistic features. ."** 13th International conference on advanced communication technology., 2011. 1620–1625.
- Gupta V., Lehal S. L., **"A Survey of Text Summarization Extractive Techniques."** Journal of emerging technologies in web intelligence, 2010: 258-268.
- Gómez-Pérez A., Fernández-López M., Corcho O., **"Ontological Engineering with examples from the areas of Knowledge Management."** e-Commerce and the Semantic Web. Springer Verlag, London, 2004.
- Haboush A., Momani A., Al-Zoubi M., Tarazi M., **"Arabic text summerization model using clustering techniques."** Journal of World Computer Science Information Technology 2(2):62–67, 2012.
- Hannah M., E. Geetha T. Mukhe S., **"Automatic Extractive Text Summarization Based on Fuzzy Logic: A Sentence Oriented Approach."** Springer-Lecture Notes in Computer Science, 2011: 530-538.
- Hatzivassiloglou V., Klavans J. L. Holcombe M. L., Barzilay R., Kan M. Y, McKeown K. R., **"SIMFINDER: A flexible clustering tool for summarization."** the NAACL Workshop on Automatic Summarization, 2001: 41-49.
- Hovy E, Lin C., **"Automated text summarization and the SUMMARIST system."** workshop on held at Baltimore. Maryland, 1998. 13-15.
- Ibrahim A, Elghazaly T., **"Rhetorical representation and vector representation in summarizing arabic text. Natural language processing and information systems."** lecture notes in computer science, vol 7934. Springer, Berlin, 2013: 421–424.
- Imam I, Hamouda A. Khalek HAA., **"An ontology-based summarization system for arabic documents (ossad)."** International Journal of Computer Application 74(17):38–43, 2013.
- Ishkewy H., Harb H. Farahat H., **"Azhar: An Arabic Lexical Ontology."** International Journal of Web & Semantic Technology (IJWeST), 2014.
- Izakian Z., Mesgari S., **"Fuzzy clustering of time series data: A particle swarm optimization approach, ."** Journal of AI and Data Mining., 2015: 39-46.
- Jaoua M, Hamadou A. B., **"Automatic Text Summarization of Scientific Articles Based on Classification of Extract's Population."** Computational Linguistics and Intelligent Text Processing. 2003. 623-634.
- Jarrar., **"The Arabic Ontology – An Arabic Wordnet with Ontologically Clean Content"**. Birzeit University, 2021.
- Jarrar M., **"Building a Formal Arabic Ontology."** the Experts Meeting on Arabic Ontologies and Semantic Networks. . Aleco Arab League. Tunis, 2011. 26-28.
- Jezek K., Steinberger J., **"Automatic text summarization, The state of the art 2007 and new challenges."** Znalosti. 2008. 1-12.
- Jin R, Abu Ata M., Xiang Y., Ruan N., **"Effective and efficient itemset pattern summarization: regression-based approaches."** 14th ACM SIGKDD international conference on Knowledge discovery and data mining., New York, NY USA, 2008. 399–407.

- John P. M., **"Mapping WordNet Instances to Wikipedia"**. Ireland: Insight Centre for Data Analytics National University of Ireland Galway, 2017.
- Jones K. S., Galliers J. R., **"Evaluating natural language processing systems: an analysis and review."** Communications of the ACM, 1996: 73-80.
- Juzaddin A., Saif O. N., **"Mapping Arabic WordNet synsets to Wikipedia articles using monolingual and bilingual features. ."** Cambridge University, 2015.
- Kaikhah K., **"Automatic Text Summarization with Neural Networks."** second international Conference on intelligent systems, IEEE. Texas , USA, 2004. 40-44.
- Kanan R.A.S.G., Jaam J., Hailat E., **"Stop-word removal algorithm for arabic language."** International Conference on Information and Communication Technologies: From Theory to Applications. IEEE. 2004.
- Kavita G., ChengXiang Z. Jiawei H., **"Opinosis: A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions."** Proceedings of the 23rd International Conference on computational Linguistics (Coling 2010). 2010. 340-348.
- Kennedy J, Eberhart R., **"Particle Swarm Optimization."** IEEE international joint conference on neural networks. 1995. 1942-1948.
- Keskes I., **"Discourse analysis of arabic documents and application to automatic summarization"**. Doctoral Thesis, University of Toulouse III-Paul Sabatier, 2015.
- Khoja, S. and Garside, R. **"Stemming Arabic Text"**. Computing Department, Lancaster University, Lancaster.,1999
- Kumar C., Pingali P. Varma V., **"Generating Personalized Summaries Using Publicly Available Web Documents. ."** ACM International Conference on Web Intelligence and Intelligent Agent Technology. DC, USA: IEEE Computer Society Washington, 2008. 103-106.
- Kumar K. V., Yadav, D., Sharma, A., **"Graph Based Technique for Hindi Text Summarization."** Information Systems Design and Intelligent Applications, 2015: 301-310.
- Kumar Y. J., Goh O. S., Basiron H., Choon, N. H., Suppiah P. C., **"A Review on Automatic Text Summarization Approaches."** Journal of Computer Science, 2016: 178-190.
- Kupiec J, Pedersen J, Chen F., **"A Trainable Document Summarizer."** SIGIR '95 18th annual international ACM SIGIR conference on Research and development in information retrieval. 1995. 68-73.
- Kyoomarsi F., Khosravi H. Eslami E. Dehkordy P. K. Tajoddin A., **"Optimizing Text Summarization Based on Fuzzy Logic."** Seventh IEEE/ACIS International Conference on Computer and Information Science. 2008. 347-352.
- Lagrin S., Redjimi M., Azizi N., **"Automatic arabic text summarization approaches."** International Journal of Computer Application 164 (5)., 2017.
- Lakshmi S., Deepthi K., Suresh C., **"Text summarization basing on font and cue-phrase feature for a single document. Emerging ICT for Bridging the Future."** 49th Annual Convention of the Computer Society of India CSI, vol. 2. 2015. 537-542.
- Lashkari M., Moattar M., **"Improved COA with Chaotic Initialization and Intelligent Migration for Data Clustering."** Journal of AI and Data Mining, vol. 5, no. 2, , 2017: 293-305.
- Lin C., **"Rouge: A package for automatic evaluation of summaries. In Text Summarization Branches Out."** the ACL-04 Workshop, vol. 8,. 2004. 74-81.4.
- Lin C., **"Training a Selection Function for Extraction."** Eighteenth Annual International ACM Conference on Information and Knowledge Management (CIKM). Kansas City, 1999. 55-62.
- Litvak M., Vanetik N., Last M., Churkin E., **"A multilingual text summarization tool."** Museec:, 2016: 73-78.
- Lloret E, Palomar M., **"Text summarisation in progress: a literature review."** Artificial Intelligence Review 37(1):1-41, 2012.
- Lloret E., Ferrández O. Muñoz R. Palomar M., **"A Text Summarization Approach Under the Influence of Textual Entailment."** 5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2008). 2008. 22-31.

- Lloret E., Palomar M., **"Tackling redundancy in text summarization through different levels of language analysis."** Comput Stand Interfaces, 2013: 35(5):507–518.
- Luhn H., **"The Automatic Creation of Literature Abstracts."** IBM Journal of Research and Development vol. 2, no. 2, 1958.
- Mann W., Thompson S., **"Rhetorical structure theory: Toward a functional theory of text organization"**. Technical, University of Southern California / Information Sciences Institute, 1988.
- Mathkour H., **"A Novel rhetorical structure approach for classifying Arabic security documents. ."** journal of Computer Theory and engineering, , 2009: 1(3), 195.
- Mazari A. Ch., Aliane H. Alimazighi Z., **"Automatic construction of ontology from Arabic texts."** ICWIT Conference on Web and Information Technologies. 2012.
- McKeown K., Robin J. Kukich K., **"Generating concise natural language summaries."** Information Processing and Management: an International Journal, 1995: 703–733.
- Meena Y., Gopalani D., **"Efficient voting-based extractive automatic text summarization using prominent feature set."** IETE J. Res. 62 (5)., 2016.
- Meena Y., Gopalani D., **"Analysis of sentence scoring methods for extractive automatic text summarization."** International Conference on Information and Communication Technology for Competitive Strategies vol. 53. 2014.
- Mirshojaei S. H., Masoomi B., **"Text Summarization Using Cuckoo Search Optimization Algorithm. ."** Journal of Computer & Robotics, vol. 8, no. 2, , 2015: 19-24.
- Moawad F., Aref M., **"A Semantic Graph Reduction Approach for Abstractive Text Summarization."** The 2012 International Conference on Computer Engineering & Systems (ICCES'2012). Cairo, Egypt, 2012. 27-29.
- Mohammed G.H., Al Zamil. Qasem Al-Radaideh., **"Automatic extraction of ontological relations from Arabic text."** Journal of King Saud University –Computer and Information, 2014.
- Mustafa M., Salah Eldeen A., Bani Ahmad S., Elfaki A., **"A comparative survey on arabic stemming: approaches and challenges. ."** Intelligence Informormation Management 9 (2), 2017.
- Natesh A., Balekuttira, S., Patil A., **"Graph Based Approach for Automatic Text Summarization."** International Journal of Advanced Research in Computer and Communication Engineering, 2016: 6-9.
- Nazari N., Mahdavi M., **"A survey on Automatic Text Summarization."** Journal of AI and Data Mining, 2019: 121-135 .
- Nenkova A., Passonneau R., **"Evaluating Content Selection in Summarization: The Pyramid Method."** the 2004 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting. 2004.
- Nenkova A., McKeown K. **"Automatic summarization. Foundations and Trends in Information Retrieval"**, 2011: 103-233.
- Nikoo M. D., Faraahi, A. Hashemi S. M. Erfani S. H., **"A Method for Text Summarization by Bacterial Foraging Optimization Algorithm. ."** IJCSI International Journal of Computer Science Issues, vol. 9, no. 4., 2012: 36-40.
- Niraj K., Kannan S. Vasudeva V., **"A graph-based unsupervised N-gram filtration technique for automatic keyphrase extraction."** International Journal of Data Mining, Modelling and Management (IJDMMM) , 2016.
- Nobata C., Sekine S., Murata, M., Uchimoto, K., Utiyama M., Isahara H., **"Sentence extraction system assembling multiple evidence. ."** Second NTCIR Workshop Meeting, . 2009.
- Oak R., **"Extractive Techniques for Automatic Document Summarization: A Survey."** International Journal of Innovative Research in Computer and Communication Engineering, 2016: 4158-4164.
- Othman B., Haggag M. Belal M., "A taxonomy for text summarization." Information Science and Technology, 2014: 43-50.

- Oufaida H, Nouali O. Blache P., **"Minimum redundancy and maximum relevance for single and multidocument arabic text summarization."** J King Saud Univ Comput Inf Sci 26(4):450–461 special Issue on Arabic NLP, 2014.
- Passino K. M., **"Biomimicry of bacterial foraging for distributed optimization and control, ."** IEEE control system magazine, vol. 22, no. 3, , 2002: 52-67.
- Patil P. D., Kulkarni N.J., **"Text Summarization Using Fuzzy Logic."** International Journal of Innovative Research in Advanced Engineering (IJRAE), 2014: Vol. 1 Issue 3.
- Pei-ying Z., , Cun-he Li., **"Automatic text summarization based on sentences clustering and extraction."** Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference. 2009.
- Pollock J., Zamora A., **"Automatic Abstracting Research at Chemical Abstracts Service."** Chemical Information and Computer Sciences, 1975: 226-232,.
- Ponzetto S. P., Navigli R., **"Knowledge-rich Word Sense Disambiguation rivaling supervised system."** 48th Annual Meeting of the Association for Computational Linguistics,. Sweden, 2010. 1522–1531.
- Prasad R., Kulkarni U., **"Implementation and evaluation of evolutionary connectionist approaches to automated text summarization."** Computer Science 6 (11), 2010: 1366–1376.
- Qaroush A., Abu Farha I., Ghanem W., Washaha M., Maali E., **"An efficient single document Arabic text summarization using a combination of statistical and semantic features."** Journal of King Saud University – Computer and Information Sciences 33, 2021: 677-692.
- Radev. D R, Hovy E, McKeown K., **"Introduction to the special issue on summarization."** Computational Linguistics (Computational Linguistics), 2002: 399–408.
- Ragunath N., Sivaranjani R., **"Ontology Based Text Document Summarization System Using Concept Terms."** ARPN Journal Of Engineering And Applied Sciences, Vol. 10, No. 6., 2015.
- Ramesh A., Srinivasa K. G. Pramod N., **"SentenceRank- A graph based approach to summarize text. ."** Application of Information and Web Technologies (ICADIWT),. Bangalore, 2014. 177-182.
- Ranjan P. A., Saha D., **"An approach to automatic text summarization using WordNet."** Advance Computing Conference (IACC),. 2014. 1169 - 1173.
- Rodríguez H., Farwell D. Farreres J. Bertran M. Alkhalifa M. Martí M., **"Arabic WordNet: Semi-automatic Extensions using Bayesian Inference."** 6th Conference on Language Resources and Evaluation LREC2008. Marrak, 2008.
- Saggion. H., Dp S., Uk E. Radev D., Teufel S., Lam W., **"Meta-evaluation of Summaries in a Cross-lingual Environment Using Content-based Metrics."** COLING-2002. 2002. 849–855.
- Salton G., **"Automatic text processing: the transformation, analysis and retrieval of information by computer"**. Addison-Wesley Longman, 1989.
- Sarda K., **"Text Summarization using Neural Networks and Rhetorical Structure Theory."** International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 6, 2015.
- Sarkar K., **"Sentence Clustering-based Summarization of Multiple Text Documents."** TECHNIA – International Journal of Computing Science and Communication Technologies, 2009: 325-335.
- Sarkar K., Nasipuri M. Ghose S., **"Machine learning based keyphrase extraction: comparing decision trees, naïve Bayes, and artificial neural networks."** Journal of Information Processing Systems, 2012: 693-712.
- Saxena S., Saxena A., **"An Efficient Method based on Lexical Chains for Automatic Text Summarization."** International Journal of Computer Applications, 2016: 47-52.
- Schwartz R. M., Imai T. Kubala F. Nguyen L. Makhoul J., **"A maximum likelihood model for topic classification of broadcast news."** Eurospeech. 1997.
- Shaalán K, Siddiqui S, Alkhatib M, Abdel Monem A. **"Challenges in Arabic natural language processing"**. Computational linguistics, speech and image processing for Arabic language. 2018. p 59–83
- Sharifi B., Inouye D., Kalita J., **"Summarization of Twitter Microblogs"**, The Computer Journal, Volume 57, Issue 3, March 2014, Pages 378–402

- Smith B. **"Ontology"**. Blackwell: Malden: Floridi (Ed.), The Blackwell Guide to Philosophy of Computing and Information. , 2004.
- Sobh I.M.A.H., **"An optimized dual classification system for Arabic extractive generic text summarization"**. Doctoral Thesis., Giza, Egypt: Faculty of engineering, Cairo university,, 2009.
- Song W., Park S., **"A Novel Document Clustering Model Based on Latent Semantic Analysis."** Third International Conference on Semantics, Knowledge and Grid (SKG 2007). China, 2007. 539-542.
- Soor S., **"Ontology Mapping"**. India: 8th Mile, Mysore Road, RVCE Post, 2015.
- Steinberger J., Jezek K., **"Evaluation measures for text summarization."** Computing and Informatics, vol. 28, no. 2,, 2009: 1001-1028.
- Steinberger J., Ježek K., **"Using latent semantic analysis in text summarization and summary evaluation."** ISIM '04. 2004. 93-100.
- Suanmali L., Naomie S. Binwahlan M., **"Fuzzy Logic Based Method for Improving Text summarization."** IJCSIS, 2009.
- Stvore K., Vanderwende L. Burges C., **"Enhancing Single-Document Summarization by Combining RankNet and Third-Party Sources."** EMNLP-CoNLL., 2007. 448-457.
- Szymanski J., Korytkowski R. Krawczyk H., **"Collaborative Approach to WordNet and Wikipedia Integration."** The Second International Conference on Advanced Collaborative Networks, Systems and Applications. 2012.
- Tanfouri I., Tlik G., Jarray F., **"An automatic arabic text summarization system based on genetic algorithms"** 5th International Conference on AI in Computational Linguistics. 2021. 195-202.
- Teufel S. Moens M., **"Sentence Extraction as a Classification Task."** ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization. 1997. 58--65.
- Thomas S., Beutenmüller C., Puente X., Remus R., Bordag S., **"Exb text summarizer."** SIGDIAL 2015 Conference. 2015. 260–269.
- Thorat B., **"Improving Text Summarization using Fuzzy Logic & Latent Semantic Analysis."** International Journal of Innovative Research in Advanced Engineering (IJIRAE), 2014: Vol. 1 Issue 4.
- Titov J., McDonald R., **"Joint Model of Text and Aspect Ratings for Sentiment Summarization."** ACL-08: HLT. 2008. 308--316.
- Turner J., Charniak E. **"Supervised and unsupervised learning for sentence compression."** 43rd Annual Meeting on Association for Computational Linguistics. Stroudsburg, 2005. 290–297.
- Turney P., **"Learning algorithms for keyphrase extraction."** Inf. Retrieval 2 (4),, 2000.
- Van Rijsbergen, C.J. **"File Organization in Library Automation Retrieval"** Journal of Documentation, Vol. 32 No. 4,, 1976: 294-317.
- Wan X., Yang J. Xiao J., **"Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction."** ACL2007. 2007.
- Yang X. S., Deb S., **"Cuckoo Search via Lévy Flights. in proceeding of World Congress ."** Nature and Biologically Inspired Computing (NaBIC 2009),. 2009. 210-214.
- Yazan T., Jaradat A., Al-Taani A., **"Hybrid-based Arabic Single-Document Text Summarization Approach Using Genetic Algorithm."** International conference on information and communication systems (ICICS). 2016.
- Yeh J., Ke H., Yang W., Meng I., **"Text summarization using a trainable summarizer and latent semantic analysis."** Information Processing and Management: an International Journal, 2005: 75–95.
- Youngjoong, K. Jungyun, S., **"An effective sentence-extraction technique using contextual information and statistical approaches for text summarization."** Pattern Recognition Letters, , 2008: 1366–1371.
- Yu B., ben Z., Xu hua Li C., **"Latent semantic analysis for text categorization using neural network."** In Knowledge-Based Systems, 900–904. 2008.

- Zadeh L., **"From Computing with Numbers to Computing with WordsFrom Manipulation of Measurements to Manipulation of Perceptions."** IEEE transaction on circuits and systems: fundamental theory and application, 1999: 105-119.
- Zajic D., Dorr B. Schwartz R., **"Automatic Headline Generation for Newspaper Stories."** ACL Workshop on Automatic Summarization/Document Understanding Conference (DUC). 2002. 78--85.
- Zhengyuan Liu., Nancy F. Chen., **"Exploiting Discourse-Level Segmentation for Extractive Summarization."** Proceedings of the 2nd Workshop on New Frontiers in Summarization, . Hong Kong, China, November 2019. 116–121.

ملاحق البحث

Appendices

الملحق (A)

قائمة الجداول

الجدول (1-9) أهم مشاريع ودراسات التلخيص الآلي للنصوص

التقنية	المشاريع والدراسات	الميزات	العوائق
التقنيات الإحصائية Statistical method	تردد الكلمات: (Luhn, 1958)، (Baxendale, 1958)، (Brandow et al., 1995)، (Kupiec et al., 1995)، (Hovy et al., 1998)، (Fachrurrozi et al., 2013) العبارات البارزة- كلمات العنوان- موقع الجملة (Baxendale, 1958)، (Edmundson, 1969)، (Brandow et al., 1995)، (Cristea et al., 2003)، (Jaoua et al., 2003)، (Wan et al., 2007) التداخل مع العنوان أو الاستعلام: (Conroy et al., 2001)، (Hovy et al., 1998).	سهولة التنفيذ لأنها فقط تستند على الخصائص الفيزيائية للنص ويمكن استخدامها لتحسين اختيار الجملة المهمة وإلغاء التكرار.	تفشل في فهم النص لأنها لا تهتم بمعنى الجمل والكلمات
تقنيات التعلم الآلي Machine learning	نايف بايس: (Kupiec et al., 1995)، (Daume et al., 2005)، (Dorr et al., 2003)، (Sarkar et al., 2012) أشجار القرار: ((Lin, 1999)، (Jin et al., 2008)، (Chen et al., 2009))) نماذج ماركوف المخفية HMM: (Schwartz et al., 1997)، (Turner et al., 2005) (Conroy et al., 2005) نماذج اللوغاريتم الخطية: ((Youngjoong et al., 2008)، (Titov et al., 2008) (Hatzivassiloglou et al., 2001) الشبكات العصبونية: (Kaikhah, 2004)، (Svore et al., 2007)، (Burges et al., (Sarda, 2015)، (2005))	تحسين الجودة على نطاق واسع من خلال إنشاء ملخصات أكثر تماسكاً وأقل تكراراً وأكثر إفادة.	تعتمد على تعقيد قواعد بيانات التدريب في الحساب، والتي تختلف من لغة إلى أخرى، وليست ثابتة من أجل المستندات كلها

		المنطق العائلي: ((Kyoomarsi et al., 2008) ، (Hannah et al., 2011) ، (Patil et al., 2014) ، (Thorat, 2014) ، (Suanmali et al., 2009))	
فهي مهمة صعبة لأنها تتطلب ذاكرة لحفظ المعلومات الدلالية مثل الأنطولوجيا وسعة المعالج بسبب المعالجة اللغوية المعقدة	ساهم استخدام هذه التقنيات في التلخيص الآلي للنصوص في تحسين الجودة على نطاق واسع من خلال إنشاء ملخصات أكثر تماسكاً وأقل تكراراً وأكثر إفادة.	التجميع: (Sarkar, 2009) ، (Anjali et al., 2013) ، (Anjali et al., 2013) ، (Bazghandi et al., 2012) ، (Pei-ying et al., 2009) ، (Gupta et al., 2010)	التقنيات الدلالية Semantic-based method
		البيان: (Kumar et al., 2015) ، (Moawad et al., 2012) ، (Kavita et al., 2010) ، (Niraj et al., 2016)	
		التحليل الدلالي الكامن LSA: (Deerwester et al., 1990) ، (Steinberger et al., 2004) ، (Brown, 2010)	
		نظرية الخطاب Discourse theory - نظرية البنية البلاغية RST: (Chengcheng, 2010) - نظرية البنية الخطابية المجزأة SDRT: (Zhengyuan et al., 2019)	
		اعتماد الأنطولوجيا: ((Brazilay et al., 2015) ، (Saxena et al., 2016) ، (Ragunath et al., 2015))	
التعقيد في القياس	تتميز هذه التقنيات بأداء جيد	حشد الجسيمات (Binwahlan et al., 2009)	تقنيات الذكاء الجماعي Swarm intelligence based method
		بحث الوقواق: (Mirshojaei et al., 2015)	
		بحث البكتريا: (Nikoo et al., 2012)	

الجدول (9-2) أهم دراسات ومشاريع تلخيص النصوص العربية

المشاريع والدراسات	التقنية
((Ferreira et al., 2012) ،(Gupta et al., 2011) ،(El-Haj et al., 2013) ، (El-Ghannam et al., 2013) ، (Alotaiby et al., 2012)) ((El-Shishtawy et al., 2012) ،(Litvak et al., 2016) ،(Abdelkrime et al., 2015)	التقنيات الإحصائية
((Belguith et al., 2014) ، (Boudabous et al., 2010) ، (El-Fishawy et al., 2013) ، (Fattah et al., 2009) ،(Sobh, 2009)) ((Tanfour et al., 2021) ،(Al Maleh et al., 2020) ،(Qaroush et al., 2021) ،(Belkebir et al., 2015) ،(Yazan et al., 2016) ((Abuobieda et al., 2012)	تقنيات التعلم الآلي
استخدام الأنطولوجيا ((AL-Khawaldeh et al., 2015) ،(Imam et al., 2013) ، (Froud et al., 2013) ،(Brazilay et al., 2015))	التقنيات الدلالية
التجميع ((El-Haj et al., 2011a) ،(Haboush et al., 2012) ،(Fejer et al., 2014)	
البيان ((Thomas et al., 2015) ،(Erkan et al., 2004)	
نظرية الخطاب ((Ibrahim et al., 2013) ،(Keskes, 2015) ،(Elghazaly et al., 2012)	
((Al-Abdallah et al., 2017) ،(AlZahrani et al., 2015))	تقنيات الذكاء الجماعي

الجدول (9-3) جمل الملخص الناتجة عن تلخيص مستند العينة باعتماد الأنطولوجيا مع قيم أوزانها ونسب خصائصها

العبارة الضعيفة	العبارة البارزة	طول الجملة	العبارة المفتاحية	التشابه مع العنوان	موقع الجملة	وزن الجملة	الجملة
1.00	0.00	0.77	0.15	0.92	2.00	3.84	لودفيج فان بيتهوفن مؤلف موسيقى ألماني ولد عام 1770 م في مدينة بون
1.00	0.00	0.27	0.14	0.08	2.00	2.49	يعتبر من أبرز عباقرة الموسيقى في جميع العصور
1.00	0.00	0.27	0.14	0.01	1.00	1.42	وأبدع أعمالاً موسيقية خالدة
1.00	0.00	0.31	0.13	0.01	1.00	1.45	له الفضل الأعظم في تطوير الموسيقى الكلاسيكية
1.00	0.00	0.69	0.11	0.03	0.00	0.83	تشمل مؤلفاته للأوركسترا تسعة سيمفونيات وخمس مقطوعات موسيقية على البيانو ومقطوعة على الكمان
1.00	0.00	0.38	0.20	0.01	0.00	0.59	كما ألف العديد من المقطوعات الموسيقية كمقدمات للأوبرا
1.00	0.00	0.58	0.27	0.01	0.00	0.86	بدأ بيتهوفن يفقد سمعه في الثلاثينيات من عمره إلا أن ذلك لم يؤثر على إنتاجه الذي ازداد في تلك الفترة وتميز بالإبداع
1.00	0.00	1.00	0.08	0.86	0.00	2.94	شهدت مدينة بون الألمانية ميلاد الفنان العبقري لودفيج فان بيتهوفن في 16 ديسمبر عام 1770
1.00	0.00	0.54	0.07	0.03	0.00	0.64	فقدت أولى أعماله وهو في الثانية عشر من عمره عام 1783 م
1.00	0.00	0.42	0.09	0.14	0.00	0.65	ثم زاد إنتاجه وذاع صيته كمؤلف موسيقى
1.00	0.00	0.65	0.06	0.01	0.00	0.72	فبالرغم من أن أباه هو معلمه الأول الذي وجه اهتمامه للموسيقى ولقته العزف على البيانو والكمان
1.00	0.00	0.15	0.50	0.00	0.00	0.65	عاصمة الموسيقى في ذلك العصر
1.00	0.00	0.54	0.07	0.03	0.00	0.64	فهل كان التأليف الموسيقي هو نوع من أنواع العلاج والتغلب على المشاكل بالنسبة لبيتهوفن
1.00	0.00	0.23	0.50	0.03	0.00	0.76	فقد أرسله حاكم بون إلى فيينا
1.00	0.00	0.08	1.00	0.00	0.00	1.08	وهناك تتلمذ على يد هايند ولكن بيتهوفن
1.00	0.00	0.42	0.73	0.03	0.00	1.18	وقد أسهمت كل هذه الدروس والاحتكاكات في تكوين شخصية بيتهوفن الفنية
1.00	0.00	0.23	0.33	0.00	0.00	0.56	غناه هو أعماله الفنية المتميزة
1.00	0.00	0.50	0.08	0.07	0.00	0.65	بدأت إصابته ببيتهوفن بصمم بسيط عام 1802
1.00	0.00	0.23	0.50	0.00	0.00	0.73	فبدأ في الانسحاب من الأوساط الفنية تدريجياً
1.00	0.00	0.62	0.06	0.00	0.00	0.68	وأمضى حياته بلا زواج يرتبط بعلاقات عدة مع سيدات صغيرات
1.00	0.00	0.15	0.50	0.00	0.00	0.65	إلا أنه لم يتوقف عن الإنتاج الفني
1.00	0.00	0.15	0.50	0.03	0.00	0.68	امتنع عن العزف في الحفلات العامة
1.00	0.00	0.31	0.38	0.00	0.00	0.69	وابتعد عن الحياة الاجتماعية واتجه للوحدة وقلت مؤلفاته
1.00	0.00	0.08	0.50	0.00	0.00	0.58	وأصبحت أكثر تعقيداً
1.00	1.00	0.50	0.08	0.04	0.00	1.62	وبالفعل مازالت أعماله حتى اليوم من أهم ما أنتجته الموسيقى الكلاسيكية العالمية
1.00	0.00	0.19	0.60	0.00	0.00	0.79	وهما السيمفونية الخامسة والتاسعة
1.00	0.00	0.19	0.40	0.00	0.00	0.59	كما أنه أحدث الكثير من التغييرات في الموسيقى

الجدول (9-4) عينات من مجموعات المفاهيم والمصطلحات المرتبطة في الأنطولوجيا العربية

المفهوم المجرد	المصطلح المجرد المرتبط بالمفهوم	تصوير شرايين
إبعاد مركب	إبعاد سيار	تضييق عصيد
	حفر هيدروليك	حال احتشال
	سبع قاطر	خناق صدر
	سيار شاحن	خناق عاد
	سيار قاطر	دوا خناق
	عرب متحرك	ديلميه دم
	مقطور	رجف بطين
	نصف مقطور	صفيح عصيد
	عناصر لانتانيديه	عصيد شريان
	غادولينيوم	أجر محاكم
الترب نادر	لانتانيديه	تدبير اصلاح
	لوتينيوم	جائح سنه
	التربي	حدث جائح
	أكسيد التريوم	محاكم أحداث
التريوم	املاح التريوم	محكم أحداث
	خواص التريوم	كتاب خواتيم
	زرنخ التريوم	نقش اختام
	فسفور التريوم	أحداث اختلاج
	فلور التريوم	اعتلال دماغ
	كبريت التريوم	امراض دماغ
	كربون التريوم	نقص كسليه
	كلوريد التريوم	يوريميه
	نترات التريوم	اختناق جئين
	هيدريد التريوم	اعط اكسجين
	يوديد التريوم	اكسجين نسب
	تايح تحويل	تنبيب رغامى
	تحكم رقم	حال اختناق
	تحويل منظوم	أخراج مجل
التماني	تصميم منظوم	اسلوب اخراج
	محاك منظوم	مخر ج صحف
	منظوم منقطع	كفاي عمل
	منظوم مستمر	مدير انتاج
	انتاج مومت	مصنع تجهيز
	تطبيق التمت	احتياط اهلاك
الامت	منظوم مومت	تحليل مال
	اثر بلتيه	تمويل قسير
	اثر طومسون	خط مال
اثر تشيرونكوف	اشعاع تشيرونكوف	عام تمويل
	تشيرنكوف	عمل جاري
	عداد تشيرونكوف	انريذالين نور انريذالين
اجاز عمل	اجاز صح	تنبيه اعصاب
	اجاز كامل	لب كظر
	عامل اجاز	نور انريذالين
اجتهاد	مجال اجتهاد	جوروس
	مصادر اجتهاد	معد انفو
	اجهاض جريم	ادمان
اجهاض	اجهاض دوا	ادمان اعط
	اجهاض عفو	اعتقاد عضو
	اجهاض متكرر	امفيتامين
	اجهاض محرض	ايتلاميد
	تجريف	ايقاف تعاط
	جريم اجهاض	تكخين
	حامل اجهاض	تعاط
	عقوب اجهاض	تعاط حشيش
	اجهز تعليق	حدوث ادمان
	تعليق مركب	حمض ايز رجيك
	تعليق مرن	سوم ادمان
	جهاز تعليق	ادنيره حديث
اجهز رصد تلك	عارض استناد	ادنيره حديق
	كراين مضاجع	ادنيره عصور
	مجال التو	ادنيره قديم
	مجموع تعليق	ادنيره مدني
	مركب اجهز	تيار وشيع
	مركب ثبات	دار قياس
	مركب مزنجر	قياس تيار
	مضاجع عجل	كهرديناميك
	نظام تعليق	مقياس غلقان
	هيدروهاي	وشيع متحرك
	هيكل مركب	اذن ظاهر
	راسد	اذن وسط
	كاسفران	اذن وظيف
	احباط ضعيف	عصب سمع
احتشال قلب امراض شرايين اكليل	سلوك عنوان	عظم ركاب
	احتشال شامل	غشا طبل
	اصاب احتشال	تغير اوستاش
	اكليل سوي	نقص سمع
	الم خناق	وسط اذن
	الم صدر	وظيف اذن
أحداث جائح	أحداث جائح	تصوير شرايين
	أحداث جائح	تضييق عصيد
	أحداث جائح	حال احتشال
	أحداث جائح	خناق صدر
	أحداث جائح	خناق عاد
	أحداث جائح	دوا خناق
	أحداث جائح	ديلميه دم
	أحداث جائح	رجف بطين
	أحداث جائح	صفيح عصيد
	أحداث جائح	عصيد شريان
اختام	اختام	أجر محاكم
	اختام	تدبير اصلاح
	اختام	جائح سنه
	اختام	حدث جائح
اختلاج	اختلاج	محاكم أحداث
	اختلاج	محكم أحداث
	اختلاج	كتاب خواتيم
	اختلاج	نقش اختام
	اختلاج	أحداث اختلاج
	اختلاج	اعتلال دماغ
اختناق	اختناق	امراض دماغ
	اختناق	نقص كسليه
	اختناق	يوريميه
	اختناق	اختناق جئين
	اختناق	اعط اكسجين
	اختناق	اكسجين نسب
أخراج صحف	أخراج صحف	تنبيب رغامى
	أخراج صحف	حال اختناق
	أخراج صحف	أخراج مجل
	أخراج صحف	اسلوب اخراج
ادار انتاج	ادار انتاج	مخر ج صحف
	ادار انتاج	كفاي عمل
	ادار انتاج	مدير انتاج
	ادار انتاج	مصنع تجهيز
ادار مال	ادار مال	احتياط اهلاك
	ادار مال	تحليل مال
	ادار مال	تمويل قسير
	ادار مال	خط مال
	ادار مال	عام تمويل
	ادار مال	عمل جاري
انريذالين	انريذالين	انريذالين نور انريذالين
	انريذالين	تنبيه اعصاب
	انريذالين	لب كظر
	انريذالين	نور انريذالين
انفو	انفو	جوروس
	انفو	معد انفو
	انفو	ادمان
	انفو	ادمان اعط
ادمان سموم	ادمان سموم	اعتقاد عضو
	ادمان سموم	امفيتامين
	ادمان سموم	ايتلاميد
	ادمان سموم	ايقاف تعاط
	ادمان سموم	تكخين
	ادمان سموم	تعاط
	ادمان سموم	تعاط حشيش
	ادمان سموم	حدوث ادمان
	ادمان سموم	حمض ايز رجيك
	ادمان سموم	سوم ادمان
	ادمان سموم	ادنيره حديث
	ادمان سموم	ادنيره حديق
ادنيره	ادنيره	ادنيره عصور
	ادنيره	ادنيره قديم
	ادنيره	ادنيره مدني
	ادنيره	تيار وشيع
ادو قياس كهرب	ادو قياس كهرب	دار قياس
	ادو قياس كهرب	قياس تيار
	ادو قياس كهرب	كهرديناميك
	ادو قياس كهرب	مقياس غلقان
	ادو قياس كهرب	وشيع متحرك
	ادو قياس كهرب	اذن ظاهر
اذن انسان	اذن انسان	اذن وسط
	اذن انسان	اذن وظيف
	اذن انسان	عصب سمع
	اذن انسان	عظم ركاب
		غشا طبل
		تغير اوستاش
		نقص سمع
		وسط اذن
		وظيف اذن

الجد(9-5) عينات من علاقات التشابه والترافق في الأنطولوجيا العربية المتولدة

نوع العلاقة	المصطلح الثاني	المصطلح الأول
تشابه	أبيا	أبأيا
تشابه	طبول	أبواق
تشابه	نفري	أبيضاض
تشابه	مصفا	أعقاب
تشابه	الانترولوجي	الانترولوجي
تشابه	مزلق	احتجاز
تشابه	فلول	أحتم
تشابه	عالي	أخمس
تشابه	نوراندريالين	اندريدالين
تشابه	بطين	أدين
تشابه	سابور	أدين
تشابه	ارثوس	ارالوس
تشابه	كاريب	اراك
تشابه	رجاز	ارجز
تشابه	راحم	ارحم
تشابه	شجاج	ارش
تشابه	اساربع	ازعج
تشابه	أخضور	آز هوري
تشابه	اشاح	اساده
تشابه	معالج	استحصار
تشابه	جحد	استحل
تشابه	بلمر	استر
تشابه	استر كان	استراخان
تشابه	استروجين	استراديول
تشابه	استريول	استراديول
تشابه	محفور	استشراق
تشابه	استدان	استقراض
تشابه	استتمساخ	استتمساخ
تشابه	مرقط	اسقمري
تشابه	ارفق	اسمع
تشابه	منقبح	اشوع
تشابه	أخرف	اصاف
تشابه	أقل	اضرب
تشابه	معار	اعار
تشابه	الثوب	أقت
تشابه	أفخاش	أفذاع
تشابه	اماجد	أكارم
تشابه	مكس	أكسال
تشابه	حاجر	أكاف
تشابه	استجار	الطاف
تشابه	البرولين	الغليس
تشابه	البروتوكاين	الغليكروتين
تشابه	المالتوز	الفركتوز
تشابه	ألهو	اللور
تشابه	عيل	أنفع
تشابه	فريخ	انيس
تشابه	المستر اول	البيسترون
تشابه	شجن	النين
تشابه	حسيره	أهات
تشابه	حصابر	اوتاد
تشابه	عتي	بستي
تشابه	ياز ١٧	بسله
تشابه	سبحل	بسمل
تشابه	نقتر	بهبان
تشابه	بيز انط	بيز انتيون
تشابه	ثاليوم	بيز موت
تشابه	توان	تابق
تشابه	تخشب	تائل
تشابه	تاريف	تاريت
تشابه	غرائبان	تابار
تشابه	كورار	تافي
تشابه	تشاكل	تياكل
تشابه	تيزير	تيزر
تشابه	استحجر	تخشب
تشابه	تسييع	تخميس
تشابه	تيرك	ترتش
تشابه	أحنيج	ترخيم
تشابه	تحمص	ترذ
تشابه	بسله	ترمس
تشابه	ديرامب	ترليم
تشابه	ستر انوسفير	تروبوسفير
تشابه	تعمير	تسييع
تشابه	مماطل	تسوييف
تشابه	شايب	تشويه
تشابه	انغمار	تشوك
تشابه	عصيد	تصلب
تشابه	تحجيب	تطعيم
تشابه	يلاعمر	تعاضد
تشابه	اكسين	تعابير
تشابه	تشفير	تعمي
تشابه	جدانوف	تغايروغ
تشابه	بلاصر	تغللات
تشابه	ميشر	تقمم
تشابه	ومضائي	تقرس

تشابه	كسل	تفبح
تشابه	تعتيق	تقص
تشابه	تطريق	تكتيل
تشابه	منقل	تكسر
تشابه	أذره	تلييد
تشابه	ثرثر	تلغو
تشابه	كادميوم	تلوريوم
تشابه	عسول	تليلات
تشابه	أجايي	تنويم
تشابه	كلكل	تنينخ
تشابه	طهرق	تهارك
تشابه	سكيل	توان
تشابه	مسود	تودع
تشابه	أعشاب	أباد
تشابه	جماع	أباد
تشابه	مركب	أبعاد
تشابه	دم	أبيضاض
تشابه	أوكالي	أبيوريماك
تشابه	هفليستون	أبيون
تشابه	براش	أثار
تشابه	موسيق	أجاء
تشابه	استراديول	أنتيل
تشابه	بروتين	أتين
تشابه	تشيرنكوف	أثر
تشابه	سطح	أثر
تشابه	فن	أثر
تشابه	هول	أثر
تشابه	عشر	أثا
تشابه	غاليينيس	أثاثر
تشابه	ألا	أجا
تشابه	عمل	أجاز
تشابه	طاعات	أجنتب
تشابه	قضا	أجهاد
تشابه	مزلق	أحجاز
تشابه	كريم	أحجاز
تشابه	جانب	أحداث
تشابه	مواد	أختيار
تشابه	لامارك	أخري
تشابه	صحف	أخراج
تشابه	فن	أخراج
تشابه	سيوت	أخوص
تشابه	سبيل	أخل
تشابه	سموم	أدمان
تشابه	سمندر	أدغفن
تشابه	انسان	أذن
تشابه	بطين	أذين
تشابه	ارلوس	ارالوس
تشابه	بنا	أساس
تشابه	بعد	استشعار
تشابه	كهز ساكن	اسقاط
تشابه	بيولوجي	اسلج
تشابه	أبوض	اسمنت
تشابه	أمنت	اسمنت
تشابه	تجار	اسناد
تشابه	عرب	اسواق
تشابه	مال	اسواق
تشابه	شاق	اشغال
تشابه	حيا	اصل
تشابه	دين	اصلاح
تشابه	زراع	اصلاح
تشابه	نقد	اصلاح
تشابه	محاكم	اصول
تشابه	مسرح	أصا
تشابه	تأهيل	أعاد
تشابه	محاكم	أعاد
تشابه	مال	أعان
تشابه	عرض	أعدن
تشابه	مستند	أعتماد
تشابه	محل	أعذار
تشابه	سحق	أعماق
تشابه	زراع	أفات
تشابه	تعليم	أقتصاد
تشابه	رفا	أقتصاد
تشابه	زراع	أقتصاد
تشابه	سكان	أقتصاد
تشابه	طاق	أقتصاد
تشابه	عمل	أقتصاد
تشابه	إذا	أقتصاد
تشابه	قياس	أقتصاد
تشابه	معرف	أقتصاد
تشابه	أفخاش	أفذاع
تشابه	طائر	أفلاغ
تشابه	هيوط	أفلاغ
تشابه	أرجاع	أكسد
تشابه	حشر	أكل
تشابه	لحوم	أكل

الملحق (B)

الإجراءات الأساسية

Main Procedures

-----حساب الخصائص وتثقيل الجمل في التلخيص باعتماد الأنطولوجيا-----

```
PROCEDURE sentence_ontology_weight (the_resno in number) IS
  CURSOR seq_cursor IS
    select distinct seq from doc_sentences where resno = the_resno order by seq;
  CURSOR term_cursor(id_1 number) IS
    select seq,synset,wrд_type,freq from doc_sentence_keyword where resno = the_resno and seq = id_1 order by freq;
  max_no number(8);
  sum_no number(8);
  word_weight number(8,2);
  seq_count number(2);
  the_count number(4);
  keyphrases_freq number(6,2);
  title_intersect number(6,2);
  the_title_sim number(6,2);
  title_sim number(6,2);
  max_length number(2);
  sen_length number(6,2);
  doc_cue_phrases number;
  seq_cue_phrases number(6);
  cue_phrases number(6);
  doc_weak_phrases number(6);
  sentence_location number(6,2);
  sentence_score number(6,2);
  the_weight number(6,2);

BEGIN
  استخراج مفهوم المصطلح--
  update doc_sentence_words w
  set synset = (select min(d.aaword) from res_synset_details where d.aaword = w.aaword and type is null) where resno = the_resno;
  update doc_sentence_words w
  set synset = (select min(r.aaword2) from res_synset_relations r where w.aaword = r.aaword1 and arel_type = الاشتمال )
  where resno = the_resno and synset is null;
  update doc_sentence_words w
  set synset = (select min(r.aaword2) from res_synset_relations r where w.aaword = r.aaword1 and arel_type = عضو )
  where resno = the_resno and synset is null;
  update doc_sentence_words w
  set synset = (select min(r.aaword2) from res_synset_relations r where w.aaword = r.aaword1 and arel_type = التضمين )
  where resno = the_resno and synset is null;
  update doc_sentence_words set synset = aaword where resno = the_resno and synset is null;
  استخراج الكلمات المفتاحية--
  insert into doc_sentence_keyword(resno,seq,wrд_type,synset,freq)
  select distinct resno,seq,wrд_type,synset,count(aaword) from doc_sentence_words where resno = the_resno group by
  resno,seq,wrд_type,synset,wrд_type;
  تثقيل المصطلحات--
  select count(seq) into seq_count from doc_sentences where resno = the_resno ;
  الجمل في الوثيقة--
  عدد الكلمات في أطول جملة--
  select max(count(synset)) into max_length from doc_sentence_keyword where resno = the_resno group by seq;
  select count(distinct synset) into doc_cue_phrases from doc_sentence_keyword k where resno = the_resno
  and exists (select 1 from doc_cue_phrases p where k.synset = p.aphrase);
  الجمل البارزة في الجمل كلها--
  حلقة الجمل--
  FOR seq_rec in seq_cursor LOOP
    أعلى تردد في الكلمة--
    select nvl(max(freq),1),count(distinct synset) into max_no,the_count from doc_sentence_keyword
    where resno = the_resno and seq_rec.seq = seq ;
    حلقة مصطلحات الجملة--
    FOR term_rec in term_cursor(seq_rec.seq) LOOP
```

```

select nvl(count(distinct seq),1) into sum_no from doc_sentence_keyword
where synset = term_rec.synset and resno = the_resno ;
--مجموع تردد المصطلح في الجمل كلها--
IF sum_no = 0 THEN
    sum_no:= 1;
END IF;
--حساب وزن المصطلح--
word_weight:=(0.01+(0.01*the_count*(term_rec.freq/nvl(max_no,1))))*
(round(log(10,(seq_count/nvl(sum_no,1))),2));
--إدخال النقل إلى جدول الكلمات المفتاحية--
update doc_sentence_keyword set weight = word_weight, all_freq = sum_no
where resno = the_resno and seq = term_rec.seq and synset = term_rec.synset ;
END LOOP;
--وسطي ثقل العبارات المفتاحية--
select avg(weight) into the_weight from doc_sentence_keyword where resno = the_resno and seq = seq_rec.seq;
--التشابه مع العنوان باستخدام علاقة التجيب--
o_seq_similar (the_resno ,seq_rec.seq,1, the_title_sim ) ;
--عدد الكلمات المشتركة مع العنوان--
select count(distinct synset) into title_intersect from doc_sentence_keyword k1 where resno = the_resno and seq = seq_rec.seq
and exists (select 1 from doc_sentence_keyword k2
where k1.resno = k2.resno and k2.resno = the_resno and k1.seq <> k2.seq and k1.synset = k2.synset and k2.seq = 1);
--عدد العبارات البارزة في الجملة--
select count(distinct synset) into seq_cue_phrases from doc_sentence_keyword k
where resno = the_resno and seq = seq_rec.seq and exists (select 1 from doc_cue_phrases p where k.synset = p.phrase);
--خاصية موقع الجملة--
IF seq_rec.seq in (1,2) THEN
    sentence_location:= 2;
ELSIF seq_rec.seq in (3,4) THEN
    sentence_location:=1;
ELSE
    sentence_location:= 0;
END IF;
--خاصية التشابه مع العنوان--
title_sim := the_title_sim * sqrt(1+title_intersect);
--خاصية طول الجملة--
IF max_length = 0 THEN
    max_length := 1;
END IF;
sen_length := the_count / max_length;
--خاصية العبارات البارزة--
IF doc_cue_phrases = 0 THEN
    oc_cue_phrases := 1;
END IF;
cue_phrases := seq_cue_phrases/doc_cue_phrases;
--خاصية الكلمات الضعيفة--
select count(distinct seq) into doc_weak_phrases from doc_sentences s
where resno = the_resno and seq =seq_rec.seq and
exists (select 1 from doc_weak_words w where s.sentence like '%||w.weak_word||"%');
IF doc_weak_phrases >=1 THEN
    doc_weak_phrases := 0 ;
ELSE
    doc_weak_phrases := 1 ;
END IF;
--ثقل الجملة--
sentence_score := doc_weak_phrases * (sentence_location + title_sim + the_weight + sen_length + cue_phrases );
update doc_sentences
set weight = sentence_score, sentence_location=sentence_location ,
title_sim = title_sim ,
keyphrases_freq = keyphrases_freq,
sen_length=sen_length,cue_phrases=cue_phrases
where resno = the_resno and seq = seq_rec.seq;
END LOOP;
insert into doc_keyword (resno,synset,wrд_type,weight)
select distinct resno,synset,wrд_type,avg(weight)
from doc_sentence_keyword where resno= the_resno group by resno,synset,wrд_type;
update doc_keyword k set weight = 1 where exists (select 1 from doc_sentence_keyword s where resno=the_resno and
k.synset = s.synset and seq in(1,2,3,4) and weight >=0.4);
commit;
END;

```

-----حساب الخصائص وتثقيل الجمل في التلخيص الإحصائي-----

PROCEDURE sentence_weight (the_resno in number)IS

CURSOR seq_cursor IS --الجمل

select distinct seq from doc_sentences where resno = the_resno order by seq;

CURSOR term_cursor(id_1 number) IS --المصطلحات

select seq,aword,wrд_type,freq from doc_sentence_keyword

where resno = the_resno and seq = id_1 order by freq;

max_no number(8);

sum_no number(8);

word_weight number(8,2);

seq_count number;

the_count number(4);

keyphrases_freq number(6,2);

title_intersect number(6,2);

the_title_sim number(6,2);

title_sim number(6,2);

max_length number(6);

sen_length number(6,2);

doc_cue_phrases number(6);

seq_cue_phrases number(6);

cue_phrases number (6);

doc_weak_phrases number (6,2);

sentence_location number(6,2);

sentence_score number(6,2);

the_weight number(6,2);

BEGIN

insert into doc_sentence_keyword(resno,seq,aword,aword1,aword2,wrд_type,freq)

select distinct resno,seq,aword,aword1,aword2,wrд_type,count(aword) from doc_sentence_words where resno = the_resno group by resno,seq, aword, aword1, aword2, wrд_type;

--تثقيل المصطلحات--

عدد الجمل في الوثيقة--

select count(seq) into seq_count from doc_sentences where resno = the_resno ;

عدد الكلمات في أطول جملة--

select max(count(aword)) into max_length from doc_sentence_keyword where resno = the_resno group by seq;

عدد العبارات البارزة في الجمل كلها--

select count(distinct aword) into doc_cue_phrases from doc_sentence_keyword k

where resno = the_resno and exists (select 1 from doc_cue_phrases p where k.aword = p.aphrase);

حلقة الجمل--

FOR seq_rec in seq_cursor LOOP

--أعلى تردد في الجملة--

select nvl(max(freq),1),count(distinct aword) into max_no,the_count from doc_sentence_keyword

where resno = the_resno and seq_rec.seq =seq ;

حلقات مصطلحات الجملة--

For term_rec in term_cursor(seq_rec.seq) LOOP

مجموع تردد المصطلح في الجمل كلها--

select nvl(count(distinct seq),1) into sum_no from doc_sentence_keyword

where aword = term_rec.aword and resno = the_resno ;

IF sum_no = 0 THEN

sum_no := 1;

END IF;

حساب ثقل المصطلح--

word_weight:=(0.01+(0.01*the_count*(term_rec.freq/nvl(max_no,1))))*
(round(log(10,(seq_count/nvl(sum_no,1))),2));

IF word_weight > 1 THEN

word_weight :=1 ;

END IF;

إدخال الثقل إلى جدول الكلمات--

update doc_sentence_keyword set weight = word_weight, all_freq = sum_no
where resno = the_resno and seq = term_rec.seq and aword = term_rec.aword ;

END LOOP;

وسطي ثقل العبارات المفتاحية--

select avg(weight) into the_weight from doc_sentence_keyword where resno = the_resno and seq = seq_rec.seq;

التشابه مع العنوان باستخدام علاقة التجيب--

seq_similar (the_resno ,seq_rec.seq,1, the_title_sim) ;

عدد الكلمات المشتركة مع العنوان--

select count(distinct aword) into title_intersect from doc_sentence_keyword k1

where resno = the_resno and seq = seq_rec.seq and

```

exists (select 1 from doc_sentence_keyword k2 where k1.resno = k2.resno and k2.resno = the_resno
and k1.seq <> k2.seq and k1.aword = k2.aword and k2.seq = 1);
عدد العبارات البارزة في الجملة.
select count(distinct aword) into seq_cue_phrases from doc_sentence_keyword k
where resno = the_resno and seq= seq_rec.seq and exists (select 1 from doc_cue_phrases p where k.aword = p.aphrase);
خاصية موقع الجملة.
IF seq_rec.seq in (1,2) THEN sentence_location:= 2;
ELSIF seq_rec.seq in (3,4) THEN sentence_location :=1;
ELSE sentence_location := 0.5;
END IF;
خاصية التشابه مع العنوان.
title_sim := the_title_sim * sqrt(1+title_intersect);
خاصية تردد العبارة المفتاحية.
IF the_count = 0 THEN
the_count := 1;
END IF;
keyphrases_freq := sum_no/the_count;
خاصية طول الجملة.
IF max_length = 0 THEN
max_length := 1;
END IF;
sen_length := the_count / max_length;
خاصية العبارات البارزة.
IF doc_cue_phrases = 0 THEN
doc_cue_phrases := 1;
END IF;
cue_phrases := seq_cue_phrases/doc_cue_phrases;
خاصية الكلمات الضعيفة.
select count(distinct seq) into doc_weak_phrases from doc_sentences s
where resno = the_resno and seq =seq_rec.seq and
exists (select 1 from doc_weak_words w where s.sentence like '%||w.weak_word||'');
IF doc_weak_phrases >=1 THEN
doc_weak_phrases := 0 ; else doc_weak_phrases := 1 ;
END IF;
ثقل الجملة.
sentence_score := doc_weak_phrases * (sentence_location + title_sim + the_weight + sen_length + cue_phrases );
إدخال ثقل الجملة.
update doc_sentences
set weight = sentence_score,
sentence_location=sentence_location ,
title_sim = title_sim ,
keyphrases_freq = keyphrases_freq,
sen_length=sen_length,
cue_phrases=cue_phrases
where resno = the_resno and seq = seq_rec.seq;
END LOOP;
insert into doc_keyword (resno,aword,wrд_type,aword1,aword2,weight)
select distinct resno,aword,wrд_type,aword1,aword2,avg(weight) from doc_sentence_keyword
where resno= the_resno group by resno,aword,wrд_type,aword1,aword2;
update doc_keyword k set weight = 1
where exists (select 1 from doc_sentence_keyword s where resno=the_resno and k.aword = s.aword and seq in(1,2,3,4)
and weight >=0.4);
commit;
END;

```

-----صياغة الملخص النهائي في التلخيص باعتماد الأنطولوجيا-----

```

PROCEDURE ontology_summary (the_resno in number , ratio in number) IS
  i number(6):=1;
  sam long;
  count_seq number(6);
  the_ratio number(6,2);
  المفاهيم التي معدل ثقل كلماتها أكبر من قيمة معينة--
  CURSOR key_word is select distinct w.word,w.aword from doc_sentence_words w,vm_doc_keywords_syn k
    where w.resno = k.resno and w.resno = the_resno and w.aword = k.synset and k.weight >=0.48;
  CURSOR sen_weight_cursor is select distinct sentence,seq,weight from doc_sentences
    where resno = the_resno order by weight desc;
  CURSOR sen_cursor is select distinct sentence,seq from doc_sentences1
    where resno = the_resno order by seq ;

  w_seq number(6,2);
BEGIN
  the_ratio := ratio;
  select count(distinct seq)*the_ratio into count_seq from doc_sentences where resno = the_resno;
  w_seq := 1;
  FOR sen in sen_weight_cursor LOOP
    insert into doc_sentences1(resno,sentence,seq,weight) values (the_resno,sen.sentence,sen.seq,sen.weight);
    w_seq := w_seq +1;
    IF w_seq > count_seq THEN
      exit;
    END IF;
  END LOOP;
  FOR i in sen_cursor LOOP
    sam := sam ||' '|| i.sentence;
  END LOOP;
  insert into doc_summary (resno,res_summary) values (the_resno,sam);
  commit;
END;

```

-----صياغة الملخص النهائي في التلخيص الإحصائي-----

```

PROCEDURE summary_sentence (the_resno in number , ratio in number)IS
  i number(4):=1;
  sam long;
  count_seq number(6);
  the_ratio number(6,2);
  w_seq number (6,2);
  cursor sen_weight_cursor is select distinct sentence,seq,weight from doc_sentences
    where resno = the_resno and weight >0 order by weight desc;
  cursor sen_cursor is select distinct sentence,seq from doc_sentences1 where resno = the_resno order by seq ;
BEGIN
  the_ratio := ratio;
  select count(distinct seq)*the_ratio into count_seq from doc_sentences where resno = the_resno;
  w_seq := 1;
  for sen in sen_weight_cursor loop
    insert into doc_sentences1(resno,sentence,seq,weight) values (the_resno,sen.sentence,sen.seq,sen.weight);
    w_seq := w_seq +1;
    if w_seq > count_seq then exit; end if;
  end loop;
  for i in sen_cursor loop
    sam := sam ||' '|| i.sentence;
  end loop;
  insert into doc_summary (resno,res_summary) values (the_resno,sam);
  commit;
END;

```

-----إجرائية النقال لتوليد الأنطولوجيا-----

PROCEDURE mapping IS

BEGIN

----- إدخال المفاهيم

```

insert into res_synsets ( word,aword,aaword,stem,pos, gloss,synsetid, source )
select distinct a.word,a.aword,b.aaword,null,'n',null,null,'ency' from res_words_titles a,res_keyword_title b
where b.title_flag = 2 and
not exists (select 1 from synsets c where b.aaword =c.aaword) and a.resno = b.resno and
a.aaword=b.aaword

```

----- إدخال مصطلحات المفاهيم

```

insert into res_synset_details (aavalue,aaword,source)
select distinct k.aaword,i.aaword , 'ency' from res_keyword_title k , res_keyword_title I
where k.weight >=0.7 and k.wrd_type = 1 and k.resno = i.resno and i.title_flag = 2 and
k.title_flag =1 and
not exists (select 1 from res_synsets t where k.aaword = t.aaword) and
not exists (select 1 from res_synset_details v where k.aaword = v.aavalue);

```

----- إدخال العلاقات

```

insert into res_synset_relations(word1,word2,relation_type,arel_type,source)
select distinct aaword1,aaword2,'co-occure','ترافق','ency'
from res_keyword_title k
where wrd_type =2 and co-occure >=0.5 and
exists (select 1 from res_synset_details s where k.aaword1 = s.aavalue ) and
exists (select 1 from res_synset_details t where k.aaword2 = t.aavalue )
insert into res_synset_relations(word1,word2,relation_type,arel_type,source)
select distinct aaword1,aaword2,'similar','تشابه','ency' from res_keyword_title k
where wrd_type =2 and similar >=0.5 and
exists (select 1 from res_synset_details s where k.aaword1 = s.aavalue ) and
exists (select 1 from res_synset_details t where k.aaword2 = t.aavalue ) ;

```

END;

----- توليد علاقة الترافق -----

PROCEDURE co_occur IS

```

CURSOR word_cursor is select distinct aaword ,aaword1,aaword2, sum(weight) freq
from res_keyword_title where wrd_type = 2 group by aaword,aaword1,aaword2 order by 1;
co_occur number(6,2);
freq_1 number(6);
freq_2 number(6);
the_spe number(2);
the_sqrt number(16,2);

```

BEGIN

```

insert into res_keywords_relations (aaword1,aaword2,sim,relation_type)
select distinct word1,word2 ,1,'co-occure' from res_keyword_title where wrd_type = 2 and weight =1 ;
FOR word_rec in word_cursor LOOP
select nvl(sum(weight),1) into freq_1 from res_keyword_title where aaword = word_rec.aaword1 and wrd_type = 1;
select nvl(sum(weight),1) into freq_2 from res_keyword_title where aaword = word_rec.aaword2 and wrd_type=1 ;
the_sqrt :=sqrt(freq_1*freq_2);
IF the_sqrt >0 THEN
co_occur := 1*((word_rec.freq)/the_sqrt);
END IF;
IF co_occur >0 THEN
insert into res_keywords_relations (aaword1,aaword2,sim,relation_type)
values (word_rec.aaword1,word_rec.aaword2,co_occur,'co-occure');
END IF;
commit;
END LOOP;

```

END;

----- توليد علاقة التشابه -----

```

PROCEDURE similar IS
    CURSOR word_cursor is
        select distinct word1 aaword1, word2 aaword2 from res_aaword1_aaword2 a
        where exists (select 1 from res_keyword_title b where a.word1 = aaword and weight >= 0.8 and speno = 1)
        and exists (select 1 from res_keyword_title c where a.word2 = c.aaword and weight >= 0.8 and speno = 1)
        order by 1,2;
    CURSOR ido_cursor(wrd_1 varchar2, wrd_2 varchar2) is
        select distinct resno from res_keyword_title a where speno=1 and
        exists (select 1 from res_keyword_title b
            where speno = 1 and a.resno = b.resno and b.aaword = wrd_1 and weight >= 0.8 ) and
        exists (select 1 from res_keyword_title c
            where speno = 1 and a.resno = c.resno and c.aaword = wrd_2 and weight >= 0.8)
        order by 1;
    sim    number(6,2);
    sum_weight number(6,2):=0;
    sum_weight1 number(6,2):=0;
    sum_weight2 number(6,2):=0;
    w_word1    number(6,2):=0;
    w_word2    number(6,2):=0;
    sum_w      number(6,2):=0;
BEGIN
    FOR term1_rec in word_cursor LOOP
        sum_weight := 0;
        select sqrt(sum(weight*weight)) into sum_weight1 from res_keyword_title where aaword = term1_rec.aaword1 ;
        select sqrt(sum(weight*weight)) into sum_weight2 from res_keyword_title where aaword = term1_rec.aaword2 ;
        FOR ido_rec in ido_cursor (term1_rec.aaword1, term1_rec.aaword2) LOOP
            select nvl(sum(weight),0) into w_word1 from res_keyword_title
                where aaword = term1_rec.aaword1 and resno = ido_rec.resno ;
            select nvl(sum(weight),0) into w_word2 from res_keyword_title
                where aaword = term1_rec.aaword2 and resno = ido_rec.resno ;
            sum_weight := sum_weight + (w_word1 * w_word2);
        END LOOP;
        sum_w := (sum_weight1*sum_weight2);
        IF sum_w <> 0 THEN
            sim := sum_weight/sum_w;
            IF sim <> 0 and term1_rec.aaword1 <> term1_rec.aaword2 THEN
                insert into res_keywords_relations (aaword1, aaword2, sim, relation_type, speno)
                    values (term1_rec.aaword1, term1_rec.aaword2, sim, 'simi', 1);
                commit;
            END IF;
        END IF;
    END LOOP;
    commit;
END;

```

ABSTRACT

Text summarization is the process of extracting or compiling important information from the original texts and presenting them in a summary form. Summarizing helps in obtaining the required information in the least time.

There has been a lot of research, studies and systems for the automatic summarization of texts in various European and other international languages, but research and studies in the field of summarizing Arabic texts are few and growing slowly (El-Haj et.al., 2011a) Most of them followed the extractive approaches and the use of statistical techniques and moved away from the abstract approaches and the use of semantic techniques based on knowledge sources such as ontology. Therefore, in this paper, we proposed an automatic summarization system for Arabic texts by adopting ontology, which is a general, monolingual and single-document extractive summarization system that uses semantic techniques based on ontology. Semantic relations between concepts derived from the Arabic ontology, which are synsets, holo_part, hyponym, and holo_member. The widely used Arabic Wordnet ontology has been used after expanding and enriching it with concepts and relationships derived from the Arabic encyclopedia using the mapping technology. The system was evaluated using Essex Arabic Abstracts Collection (EASC) (El-Haj et. al., 2010) , and measures of Precision, Recall and F measure. The results showed that the average precision, recall, and F measure in summarizing using the proposed ontology are good proportions compared with the proportions of automated summarization systems for Arabic texts and with the summarization systems in general.

Keywords: Automatic Sumarization, Arabic text summarization, Extractive Summarization, Abstractive Summarization, Semantic Summarization, Arabic Wordnet, Ontology, Arabic Ontology , Mappin Arabic Encyclopedia.



Syrian Arab Republic
Damascus University
Mechanical & Electrical Engineering Faculty
Department of Computer Engineering & Automation

Automatic Summarization of Arabic Text Using Ontology

**A dissertation Submitted in fulfilment of the
requirements for the degree of (PhD)
In computer Engineering and Networkes**

by
Eng. Sundos M. Said Al-Halabi

Supervisor
Dr. Mohamad Mazen Almahayri

2022-2023