نظام تلخيص آلي للنصوص العربية

د.يسرى مالك ضمد
مدرس
جامعة البصرة |كلية العلوم | الحاسبات
Yus_malik@yahoo. com

سهاد مهجر كريم
جامعة البصرة| كلية العلوم | الحاسبات
suhad_althaher83@yahoo. Com

المستخلص

يلعب التلخيص الآلي دورا مهما في ظل تطور الشبكة العنكبوتية والذي يمثل أهم الطرق لإدارة الكمية الكبيرة لمعلومات النص التي يحتاجها المستخدم للقراءة وذلك بتقليل كميته ليساعده بتحديد فيما إذا كان له علاقة بالمعلومات التي يحتاجها ام لا.

في هذا البحث قدمنا نظام تلخيص آلي للنص العربي محاولين بذلك تقليد سلوك الإنسان في كتابة الخلاصة للنص المقدم. يتناول بحثنا طريقة هجينة (إحصائية ، لغوية) حيث اقترحنا عدة معادلات رياضية بدمج مجموعة من خصائص النص لمعالجة النص الأصلي والحصول على خلاصة متماسكة تحتوي على المعلومات المفيدة بعدها تم اختيار أكفأها في إنتاج خلاصة جيدة وباستخدام خوارزمية MMR تم إزالة الفائضية من جمل الخلاصة الناتجة. ثم قيمت الخلاصة النهائية بمقارنتها مع الخلاصات التي كتبها خبراء في اللغة العربية باستخدام نظام التقييم ROUGE المتمثل بقياس نسبة الدقة Precision والاسترجاع Recall ومقياس F-Measure. كانت النتائج جيدة ومشجعة جدا كما موضحة في نتائج التقييم علما أن أنظمة التلخيص للغة العربية قليلة جدا مقارنة مع بقية اللغات. تم برمجة النظام بلغة دلفي 7.

الكلمات المفتاحية

التلخيص الآلي للنصوص، خوارزمية MMR، F-Measure.

1 - المقدمة Introduction

أصبحت معالجة اللغات الطبيعية (NLP) حقل ضروري ومهم في مجال علم الحاسوب في عصرنا الراهن حيث أن التطور في تطبيقاتها اكتسب أهمية عالية في عصر الشبكة العنكبوتية (World Wide Web) بسبب الزيادة المستمرة بكمية المعلومات المتوفرة على تلك الشبكة حيث المستخدمون يبحثون عنها مما تطلب الحاجة إلى تطبيقات جديدة تجسدت بظهور تقنيات جديدة للسيطرة على تلك المعلومات ، ولتسهيل إدارتها. احد التطبيقات المهمة لمعالجة اللغات الطبيعية هي تلخيص النصوص Text Summarization. **Waleed2005 ,Manel et.al (2008)** ولهذا يعد التلخيص الآلي على انه ضرورة من ضروريات الحياة التي يحتاجها الإنسان سعيا إلى الاقتصاد في الجهد وتوفير الوقت في ظل تطور الثورة المعلوماتية التي يحتاجها في وقتنا المعاصر.

يعرف التلخيص الآلي للنصوص على انه إنتاج تمثيل اقصر للمعلومات المهمة من نص واحد أو أكثر مع الحفاظ على الفكرة الأساسية للنص الأصلي آليا.الهدف من التلخيص

الآلي هو اخذ المعلومات المهمة من المصدر وتمثيلها وتقديمها إلى المستخدم بشكل مختصر وبأسلوب يناسب احتياجاته **(Eduard 2003)**.

1 - **أنواع أنظمة التلخيص الآلي (Jum & Lin (2009, Elena & Manual 2011**

- **نظام تلخيص نص واحد (single document system) أو المتعدد النصوص (multi-document system)**

يعتمد على عدد النصوص المدخلة إلى النظام ، إذا كان الإدخال هو نص واحد فان نظام التلخيص يطلق عليه نظام تلخيص نص واحد أما إذا كان الإدخال أكثر من نص فان نظام التلخيص يطلق عليه نظام تلخيص متعدد النصوص.

- **نظام ينتج خلاصات دلالية (indicative summary) وخلاصات غنية بالمعلومات (informative summary)**

تعرف الخلاصات الدلالي ة بأنها خلاصات تعطي معلومات مختصرة عن المواضيع

الأساسية للنص بحيث يتم الإبقاء على المعلومات الأكثر أهمية في المصدر أما الخلاصات

الغنية بالمعلومات تعرف على أنها خلاصات تحتفظ بالتفاصيل المهمة من النص

المصدر.

- **نظام تلخيص استخلاصي (Extract) ونظام تلخيص تجريدي(Abstract)**

التلخيص الاسخلاصي يعتمد على طرق الاستخلاص لإنتاج خلاصات استخلاصية باختيار الأجزاء المهمة من النص بينما التلخيص التجريدي لإنتاج خلاصات تجريدية من خلال إعادة صياغة لوحدات النص بالاعتماد على آليات فهم اللغة الطبيعية.

- **أنظمة تعتمد على لغة واحدة (monolingual)وأنظمة تلخيص تعتمد على أكثر من لغة (multilingual)**

الخلاصات النصية يكون مصدرها نص أو أكثر من نص تكتب في لغة واحدة أو أكثر

من لغة.

- **نوع المصدر (genere)**

تختلف أنواع الادخالات إلى التلخيص الآلي فقد يكون نصا أو معلومات الوسائط المتعددة (multimedia) مثل الصور والفيديو وتسجيلات الصوت

- **المجال (domain)**

نوع المستندات قيد التلخيص قد تكون مقالات إخبارية ، بحوث علمية ، نصوص قانونية وطبية.

- **المحتوى الناتج من الخلاصة (content)**

تختلف الخلاصة من حيث المحتوى إلى خلاصة عامة و خلاصة تعتمد على استفسا ر فالخلاصة العامة **(generic summary)**هي خلاصة تعطي تركيز متساوي إلى المعلومات المختلفة التي يحتويها النص وتعطي تغطية متوازنة لكل محتوياته ، أما الخلاصة التي أساسها الاستفسار **(query based summary)** فهي خلاصة تعطي تركيز فقط على المعلومات التي يحتاجها المستخدم أو ممكن اعتبارها إجابة لسؤال المستخدم وفي هذه الحالة الإدخال إلى نظام التلخيص استفسار يضعه المستخد م إلى جانب النص أو النصوص المراد تلخيصها.

2 - **الأعمال السابقة Related Work**

ظهرت فكرة التلخيص الآلي في نهاية 1950. بدأ الباحث Luhns أبحاثه في التلخيص عام 1959 بالاعتماد على مبدأ التكرار(Term Frequency) لكلمات النص وفي عام 1961 قدم الباحث Edmundson خلاصته بالاعتماد خاصية الموقع(Location Feature) لجمل النص واستمر بتطوير أبحاثه في هذا المجال حتى عام 1969**(Horacio 2008)**. بعد ذلك ظهرت العديد من التقنيات منها : في عام 1995 اقترح الباحث Kupiek خوارزمية تعلم للحصول على الخلاصة باستخدام مصنف بيزين (Bayesian

Classifier) لحساب احتمالية لكل جملة **(Julian et.al 1995)**، وظهر مفهوم استخدام العلاقات البلاغية (**Rhetorical Structure Theory**(RST)) كأحد الطرق لترابط وحدات النص مع بعضها لتوليد الخلاصة من النص المدخل بعد تمثيله في بنية هرمية وهي خوارزمية اقترحها الباحث Marcu عام 1999 وتعرف (RST) بأنها عدد من العلاقات البلاغية التي تربط وحدات النص لتوليد شجرة إعراب للنص ثم تطبي ق خوارزميات على الشجرة لاختيار الوحدات المهمة منها إلى الخلاصة **(Waleed 2005)**. و وفي عام 2001 اقترح الباحثان lin&gong الطريقة الجبرية في التلخيص (Semantic (Latent Analysis)لاستخلاص المعنى من الكلمات ومن الجمل باستخدام نموذج (Singular Vector Decomposition(SVD)) لإيجاد العلاقات بين الجمل. ثم ظهرت تقنية جديدة في التلخيص باعتماد الرسم البياني للنص (graph-based) التي اقترحها الباحثين Radav & Mihalca في عام 2004 بتمثيل النص بهيكل بياني عقده (nodes) هي الجمل وحوافه (edge) هي التشابهات بين الجمل**(Rada 2008)**. ومن الخوارزميات الشائعة لإزالة الفائضية من النص هي خوارزمية MMR الشائعة لتلخيص النصوص**(Carbonell & Goldstien 1998).**

3 - النظام المقترح Proposed System

الطريقة المقترحة للتلخيص الآلي للنص العربي في هذا البحث تتألف من الربط بين الطريقة الإحصائية والطريقة اللغوية من اجل الحصول على خلاصة جيدة ومتماسكة ويتكون من عدد من المراحل المتتالية لمعالج ة النص وإنتاج الخلاصة والشكل رقم (1) يوضح مخطط عام لعمل النظام.

4-1مرحلة المعالجة الأولية (preprocessing)

في هذه المرحلة ، نظامنا يعتمد على إدخال النص الواحد وتجري عليه عدة معالجات والمتمثلة في :

- معالجة التقطيع المتمثلة بتقطيع النص إلى عدد من الجمل التي تحدد بواسطة النقطة (.) وتقطيع الجمل إلى كلمات بواسطة الفراغ.
- معالجة إيجاد الجذع (stemming)على كلمات النص للحصول على اصل الكلمة وهي معالجة صرفية وذلك بالاعتماد على المعجم الساند للنظام للحصول على اصل الكلمة بإزالة السوابق واللواحق منها.
- تحديد أصناف كلمات النص هل هي اسم ، فعل ، صفة أو حرف.
- استبعاد الكلمات الأكثر ترددا من الحسابات في استرجاع المعلومات من النص مثل (هي ، هو ، هؤلاء ، بالنسبة ، بالإضافة ، لكن .....الخ) والتي يطلق عليها كلمات التوقف (stop of word) . ناتج هذه المرحلة متجهان هما:

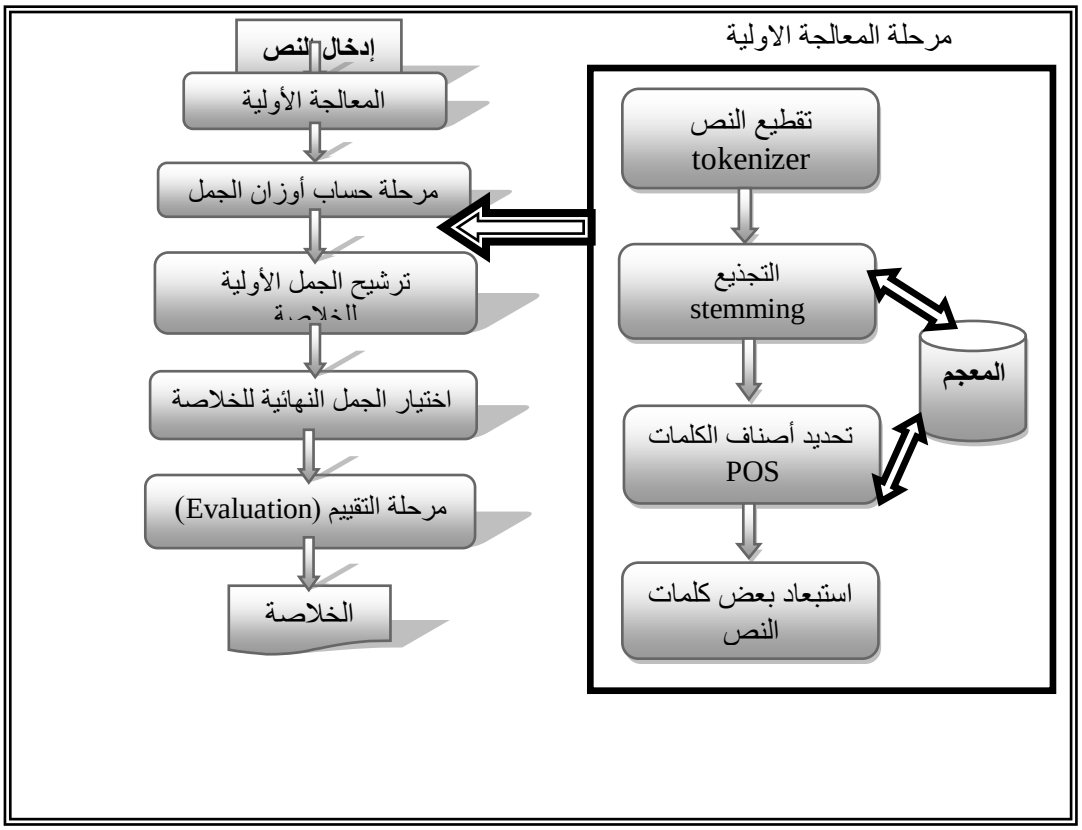
حيث أن :-

n هي العدد الكلي لكلمات النص

m هي العدد الكلي لجمل النص

$$D=\{W1, W2, \ldots Wn\}$$

$$D=\{S1, S2, \ldots Sm\}$$



4

-2مهمة حساب الأوزان (compute sentence weight)

الجملة هي المكون الأساسي للنص في جميع اللغات ومنها اللغة العربية وهي لغة البحث حيث تبنى من الكلمات والكلمة هي اقل وحدة دلالية تشير إلى المعنى. وتتضمن المهمة معالجتين هما:

1 - إعادة تركيب الجملة بعد تجذيع كلماتها لان العمليات الحسابية تجري على اصل الكلمة (الجذع) للحصول على نتائج جيدة في استرجاع المعلومات الصحيحة من النص.

2 - حساب الأوزان ، يتم حساب وزن لكل جملة من جمل النص. وزن الجملة يحسب بالاعتماد على عدد من خصائص النص، الخصائص التي استخدمناها في بحثنا الحالي هي:

1 - تكرار الكلمة (frequency word) :- في هذه المعالجة يحسب التكرار للكلمات الجذع في النص ويعد أهم وأقدم المقاييس الشائعة والمؤثرة في التلخيص الآلي للنصوص باعتبار أن الكلمات الأكثر تكرارا في النص تحمل المفاهيم الأساسية له فيحسب عدد مرات حدوث الكلمة في النص المدخل والكلمات التي تحصل على قيمة تكرار أعلى من غيرها تكون أكثر أهمية. تكرار الكلمات يحسب من المعادلة الآتية:

$$TF(W)=\sum_{J=1}^{n} Wi \quad \text{.....(1}$$

حيث ..ي ب.

2 - موقع الجملة (sentence position):- هو التسلسل الذي تظهر به في النص فمن ترتيب الجملة في النص يمكن تحديد أهميتها.

3 - طول الجملة (sentence length):- طول الجملة يمثل عدد كلماتها بعد حذف كلمات التوقف منها.

4 - عدد الأفعال (no. verb):- يعتبر الفعل مركز ثقل الجملة من حيث المعنى والتركيب فيتم ايجاد عدد الأفعال الموجودة في الجملة.

5 - التشابه مع العنوان (similarity to title):- النص وفي أي لغة ومنها اللغة العربية يتكون من جزئيين ، الأول عنوان النص والثاني محتوى النص ، ويعرف العنوان على انه

يحمل الأفكار الرئيسية التي يدور حولها محتوى النص لذلك يمكن الاستفادة من تشابه جمل النص كل على حدة مع عنوانه لإسناد وزن كل جملة وبالتالي يحدد أهميتها.

6 - تكرار الكلمات البارزة :- في كل نص يوجد عدد من الكلمات التي تحمل فكرة النص ومن ميزات هذه الكلمات أنها تحمل أعلى تردد من غيرها ، فتم جمع هذه الكلمات والتي أطلقنا عليها (الكلمات البارزة) في قائمة واستعمالها كمقياس لأهمية الجملة التي تحويها ، فكل جملة قد تحتوي على صفر أو أكثر من الكلمات البارزة وعلى هذا الأساس يتم إسناد وزن لهذه الجملة باستخدام هذه الخاصية.

تم دمج الخصائص أعلاه في عدد من المعادلات الرياضية المقترحة للحصول على وزن كل جملة. ثم اختيرت المعادلة التي اعطت نتائج جيدة مقارنة مع الاخريات.

$$\mathbf{Score(S) = S}_{\text{(Frequency)}} \mathbf{+ S}_{\text{(Length)}} \mathbf{+ S}_{\text{(Position)}} \mathbf{+ S}_{\text{(similarity)}} \mathbf{+ S}_{\text{(gist frequency)}} \mathbf{+ S}_{\text{no. verb}} \quad \text{.....(2}$$

حيث أن:

S (Frequency) هي المجموع الكلي لوزن تكرار الكلمات في الجملة حسب المعادلة أدناه:

$$\mathbf{S(Frequency)} = \sum_{i=1}^{n} \sum_{j=1}^{m} (1+\log(\text{frequency}(w))) \quad \text{.....(3}$$

حيث ان:

n هي العدد الكلي لكلمات الجملة قيد المعالجة ، m هي العدد الكلي لتكرار كلمات النص.

S (length) يحسب وزن طول الجملة كالأتي:

$$\mathbf{S(length)} = \sum_{i=1}^{n} W \quad \text{.....(4}$$

حيث أن:

n هي عدد كلمات الجملة بعد حذف كلمات التوقف

S (position) يحسب وزن موقع الجملة كالأتي:

$$\mathbf{S\ (position)} = 1/\text{position(s)} \quad \text{.....(5}$$

S (similarity) يحسب وزن تشابه جمل النص مع العنوان باستخدام مقياس تشابه الكلمات المشتركة (common word) كالأتي:

$$\mathbf{S(similarity)} = \sum_{i=1}^{n} SIM(Si , Sj) \quad i>=j \quad \text{.....(6}$$

S (gist frequency) يحسب وزن الجملة بحساب تكرار الكلمات البارزة الموجودة فيها و كالأتي :-

$$S(\textbf{Frequency}) = \sum_{i=1}^{n} \sum_{j=1}^{m} (\text{frequency(gist word)}) \quad \textbf{.....(7}$$

3-4مرحلة ترشيح الجمل الأولية للخلاصة (sentence filtering)

بعد تحديد أوزان الجمل ، يتم اختيار الجمل المهمة منها بعد ترتيبها تنازليا وبالاعتماد على نسبة الضغط (compression rate) المتمثلة بعدد جمل الخلاصة وهي نسبة جمل النص الأصلي إلى نسبة التلخيص المعطاة والتي تحسب من المعادلة الآتية:

Compression rate = (Document length * summarization rate) /100 **.....(8**

استخدمنا نسبتين للتلخيص في البحث هما 25% و40% .

-4مرحلة اختيار الجمل النهائية للخلاصة

عند اختيار الجمل بأعلى الأوزان من المرحلة السابقة ووضعها في الخلاصة فان الخلاصة قد تحتوي على معلومات متكررة لذلك يتم اختيار بعض الجمل لإزالة الفائضية منها باستخدام خوارزمية MMR **(Kamal 2009, Rasim et.al 2007)** والتي تعمل كالأتي:

1- Sort the sentences in decreasing order of their scores .
2- Select the top ranked sentence first .
3- Select the next sentence from the ordered list and include into the summary if this sentence is sufficiently dissimilar to all of the previously selected sentences , the similarity is computed from cosine similarity measure defined in the following formula :-

$$Cos(Si, Sj) = \frac{\sum_{j=1}^{m} wij \times wlj}{\sqrt{\sum_{j=1}^{m} wij^2 + \sum_{j=1}^{m} wlj^2}}$$

4- Continue selecting sentences one by one until the predefined summary length is reached .

اعتمدنا مقياسين للتشابه هما:

1 - مقياس الجيب تمام :- في مقياس ال(Cosine Similarity) يتم تمثيل الجملتين في متجه ال (VSM) لعمل تشابه بينهما بالاعتماد على المعادلة السابقة التي ذكرناها في الخوارزمية رقم (1).

2 - مقياس الكلمات المتشابهة :- أيضا في هذا المقياس يتم فيه تمثيل النص بنموذج (VSM) ويحسب التشابه من المعادلة الآتية:

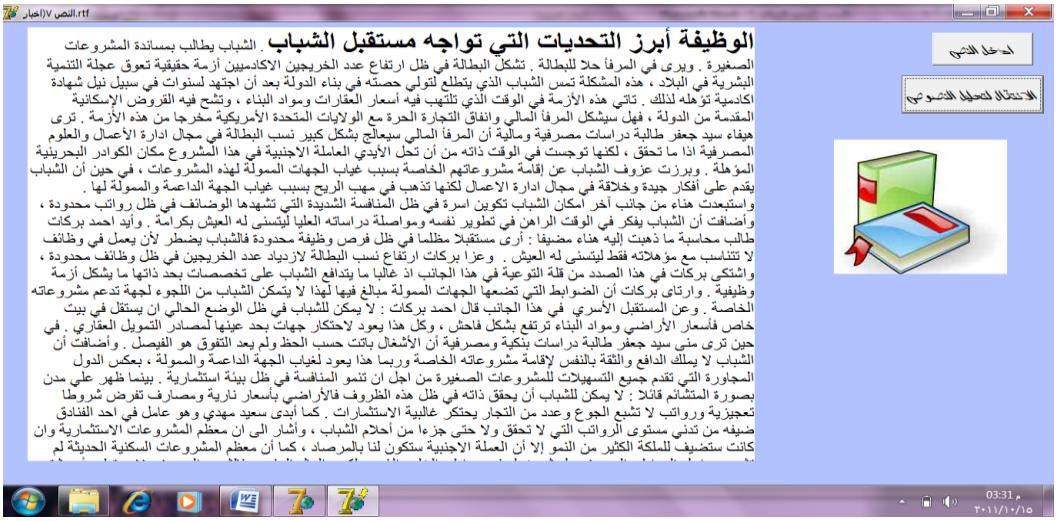
$$S(similarity) = \sum_{i=1}^{n} SIM(Si , Sj) \quad i>=j \quad .....(9$$

بعد عمل تشابه للجمل باستخدام هذين المقياسين كلا على حدة إذا كانت نسبة التشابه بين الجملتين اقل من 0.7 يتم اختيار الجملة.

وعند استخدامنا الخوارزمية رقم (1) عرضنا عليها المقياسين أعلاه في كل دورة تنفيذ لملاحظة تأثير كل مقياس على الخوارزمية في إزالة الفائضية من جمل الخلاصة كما سنوضحه في عرض نتائج التقييم كفرق بين المقياسين.

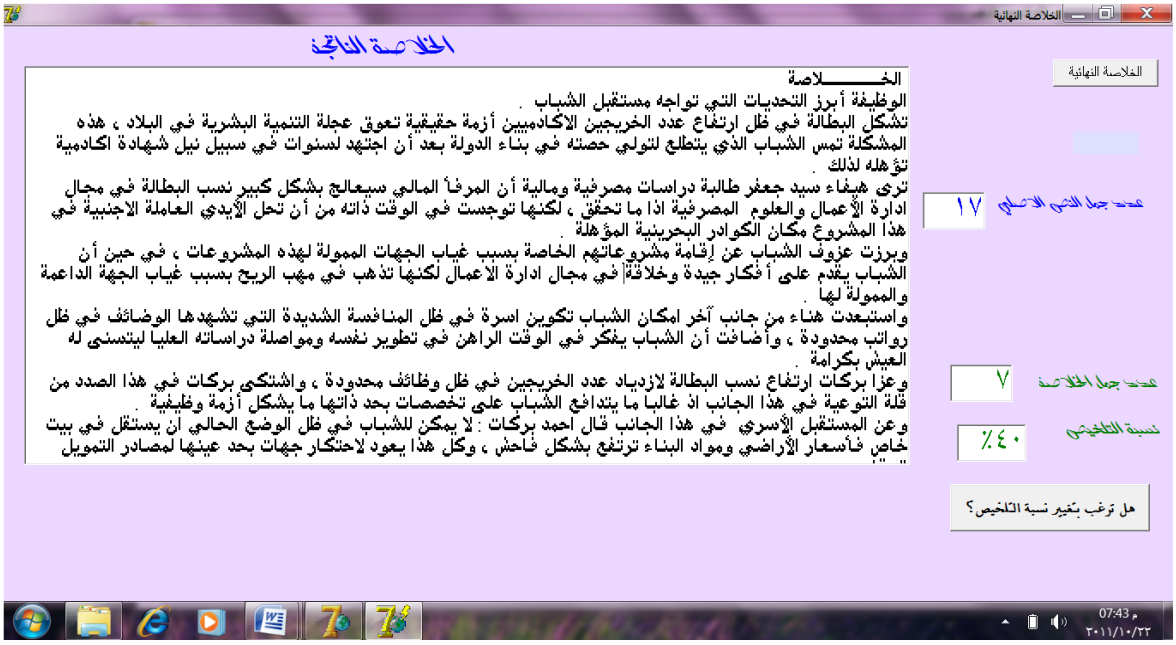
وبعد اختيار الجمل من الخوارزمية أعلاه توضع في الخلاصة بالترتيب الذي ظهرت فيه تلك الجمل في النص الأصلي لمساعدة القارئ في الحصول على خلاصة متماسكة ومتسلسلة.

4 - مثال النص المدخل والخلاصة الناتجة منه



شكل رقم (2) نافذة إدخال النص

شكل رقم (2) نافذة الخلاصة الناتجة بنسبة تلخيص 25%



شكل رقم (3) نافذة الخلاصة الناتجة بنسبة تلخيص 40%

5 - مقاييس التقييم

يعتبر تقيم أنظمة التلخيص احد التحديات ومهمة ضرورية وحاسمة في مجال التلخيص الآلي للنصوص وذلك لتحديد انجازية الطرائق والتقنيات المستخدمة في التلخيص(Dalian et.al 2002).

طرائق تقييم التلخيص الآلي يمكن أن تصنف بشكل واسع إلى صنفين(Kaustubh at.el 2007, Mani at.el 1999) هما:

أ - **مقياس التقييم الجوهري (intrinsic)** وهو مقياس تقييم للخلاصة يتم فيها تقدير جودة التلخيص مباشرة من تحليل الخلاصة من ناحية تغطية الأفكار الرئيسية عند مقارنتها مع الخلاصة المثالية التي يكتبها الخبير اللغوي.

ب - **مقياس التقييم العرضي (extrinsic)** وهو مقياس تقييم الخلاصة يتم فيها تقدير جودة التلخيص وذلك بالاعتماد على كيفية تأثيرها في إكمال المهام الأخرى كاسترجاع المعلومات من الخلاصة مقارنة بالنص الأصلي.

سنعتمد في هذا البحث على التقييم الجوهري باستخدام اشهر مقاييس التقييم وهو نظام ROUGE الذي اقترحه الباحثان Lin & Hovy في عام 2003 وهو اختصار ل Recall-Oriented Understudy Evaluation) Gisting) ويعتبر معيار لتقييم انجازية أنظمة التلخيص ، وهذا المقياس ينتج قيمة عددية يمكن استخدامها كمقارنة مع خلاصات مختلفة لنفس النص المدخل بحساب عدد التداخلات بين الخلاصة التي ولدها النظام مع الخلاصة التي كتبها الخبير باستخدام الدقة (Precision) والاستدعاء(Recall) و مقياس F-Measure)**Yihong & Xin (2001**:

$$\text{Precision}(R) = \frac{S_{man} \cap S_{auto}}{S_{auto}} \quad \text{..... (10}$$

$$\text{Recall}(R) = \frac{S_{man} \cap S_{auto}}{S_{man}} \quad \text{..... (11}$$

حيث S man تمثل الجمل التي اختارها الخبير اللغوي ، S auto تمثل الجمل التي اختارها النظام
S man ∩ S auto تمثل الجمل المشتركة بين النظام والخبير.

$$F\text{-measure} = \frac{2 * P * R}{P + R} \quad \text{..... (12}$$

6 - نتائج التقييم

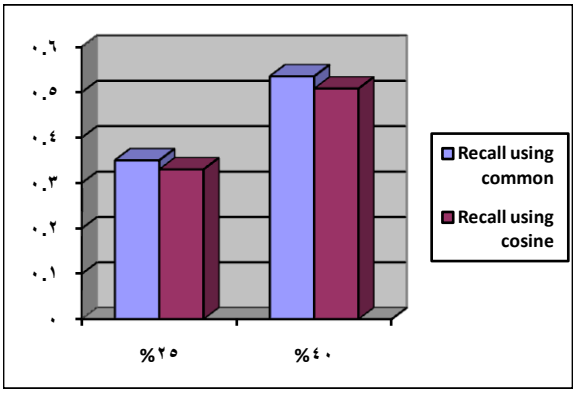
لفحص انجازية نظامنا الحالي ، عرضنا مجموعة من النصوص العربية المتكونة من 25 نصا تنوعت بين نصوص إخبارية سياسية ، اقتصادية ورياضية على خبيرين لغويين في مجال اللغة العربية ، كل منهما قدم خلاصة مختلفة عن الأخر وهذا أمر طبيعي لان لكل شخص أسلوبا ما في كتابة الخلاصة ، محاولين بذلك الاستفادة من خلاصات كلا الخبيرين لفحص انجازية نظامنا الحالي في استرجاع المعلومات من النص الأصلي بالمقارنة مع خلاصاتهم.

اعتمدنا على نسبتين للتلخيص هما 25% و40% من النص الأصلي فكانت النتائج بعد حساب المعدل لنتائج 25 نصا التي عرضت على النظام كما موضحة في الجدولان (1) و(2).

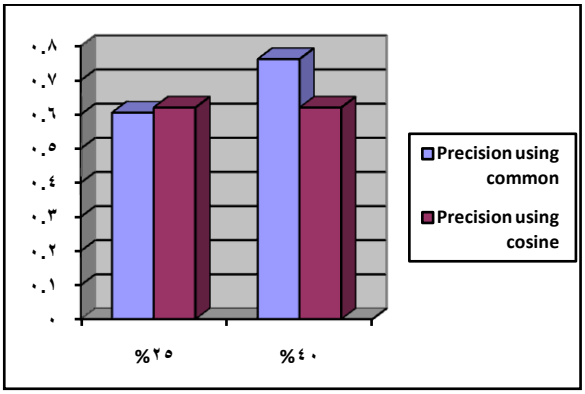
الجدولان (1) و (2) يعرضان نتائج الدقة (Precision) الاسترجاع (Recall) ومقياس (F-Measure) التي تم الحصول عليها عند عمل مقارنة بين الخلاصات المولدة آليا مع الخلاصات التي كتبها الخبيرين.

جدول (1) النتائج عند المقارنة مع خلاصات الخبير الأول

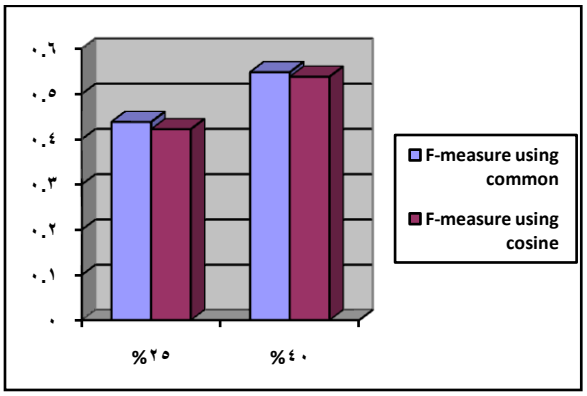
المعادلة المقترحة	نسبة التلخيص	مقياس التشابه	Precision	Recall	F-Measure
المعادلة المقترحة	%25	Cosine similarity	0.620	0.330	0.422
		Common word	0.605	.0350	0.438
	%40	Cosine similarity	0.620	0.508	0.538
		Common word	0.762	.535	0.548



شكل رقم(4) حساب الاسترجاع باستخدام مقياسين للتشابه للخبير الأول



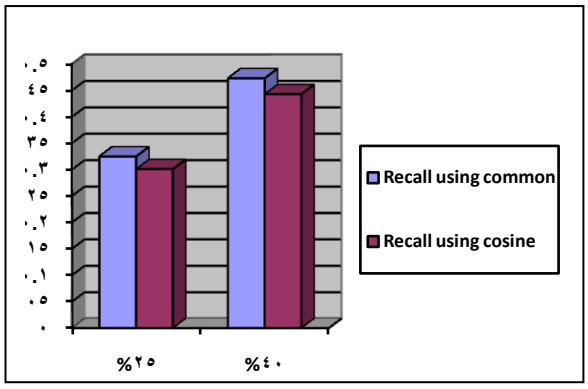
شكل رقم (5) حساب الدقة باستخدام مقياسين للتشابه للخبير الأول



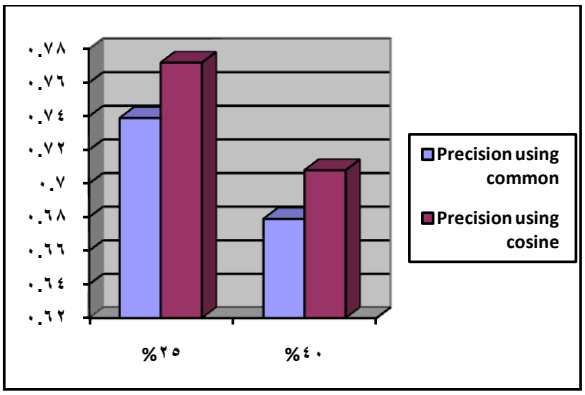
شكل رقم (6) حساب F-measure باستخدام مقياسين للتشابه للخبير الأول

جدول (2) النتائج عند المقارنة مع خلاصات الخبير الثاني

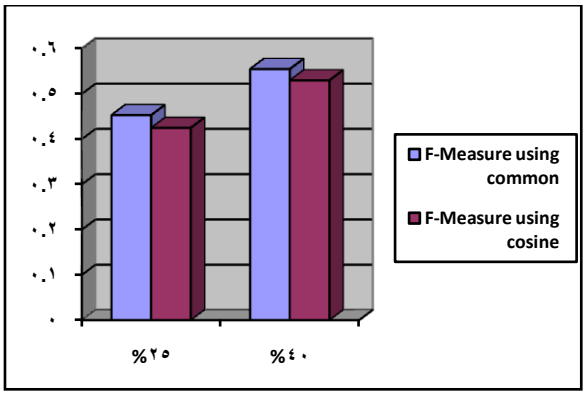
المعادلات المقترحة	نسبة التلخيص	مقياس التشابه	Precision	Recall	F-Measure
المعادلة المقترحة	%25	Cosine similarity	0.772	0.301	0.425
		Common word	0.739	0.325	0.453
	%40	Cosine similarity	0.708	0.443	0.530
		Common word	0.679	0.437	0.555



شكل رقم (7) حساب الاسترجاع باستخدام مقياسين للتشابه للخبير الثاني



شكل رقم (8) حساب الدقة باستخدام مقياسين للتشابه للخبير الثاني



شكل رقم (9) حساب F-measure باستخدام مقياسين للتشابه للخبير الثاني

7 - الاستنتاجات Conclusion

الطريقة التي عرضناها لتلخيص النص العربي بالاعتماد على مزج العديد من خصائص النص ، حصلنا العديد من الاستنتاجات منها :

- من خلال معادلاتنا الرياضية المقترحة في اختيار الجمل المرشحة للخلاصة استطعنا أن نحصل على تقييم عالي وجيد كما وضحنا في نتائج التقييم فكانت الأقرب إلى خلاصة البشر.
- كانت الخلاصة الناتجة عامة حيث يمكن استخدام النظام في أي مجال للتلخيص فقط بتزويد المعجم الساند للنظام بالمفردات اللغوية الخاصة بذلك المجال.
- كان لخاصية الكلمات البارزة في المعادلة المقترحة تأثير في إنتاج خلاصة جيدة مقارنة مع عدم وجودها فيها، بالإضافة إلى الخصائص الأخرى كان لكل منها تأثير متفاوت.
- باستخدامنا مقياسين للتشابه بين الجمل ، لاحظنا بان مقياس الكلمات المشتركة أعطى نتائج أفضل من مقياس الجيب تمام وهذا كان كاقتراح لتعديل خوارزمية MMR .

8 - المصادر Reference

Carbonell & Goldstein. "**The use of MMR,diversity-based re-ranking for reordering documents and producing summaries"**. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne,Australia, pages 335–336, 1998.

Kaustubh Patil , Pavel Brazdil . " **Sumgraph : text summarization using centrality the pathfinder network** ".International Journal on computer service and information systems, pages 18–32, 2007.

Manel Ben Abdallah , Chafik Aloulou & Lamia Belguith. " **Toward A platform for Arabic automatic summarization**" .the international Arab Conference on information technology (ACIT) , Hammamet ,December 16-18 , 2008 .

Rada Mihalcea. "**Graph-based Ranking algorthims for sentence extraction applied to text summarization**". published in proceedings of ACL , 2008 .

Rasim M. Algulev & Ramiz. "**experimental investigating the F-Measure as similarity measure for automatic text summarization**". Published in institute of information technology of Azerbaijan national academy of sience , pages 278-287 , 2007.

Waleed AL Sanie. " **Toward on infrastructure for Arabic text summarization**". Master thesis. Kingdom of Saudi Arabia. 2005.

Yihong Gong & Xin Liu. "**Generic text summarization using relevance and latent semantic analysis** ". published by ACM pages(19-25). 2001.

Dalians, H. ,Hassel, M. ,Wedekind. **"From SewSum to ScandSum : Automatic text summarization for the Scandinavian languages"**. In Proceedings of Holmboe , H(ed),Nordisk sprogteknologi ,2002.

Eduard hovy. ***"text summarization".*** the oxford handbook of computational linguistics oxford . oxford university press .2003.

Elena Eloret & manual palomar. "**text summarization in progress a literature** ". published by Springer science .2011.

Horacio saggion. **"A robust and Adaptable summarization tool**". traitement automatique des languages volume 49-n pages(103-125) .2008.

Inderjeet mani, David House ,David Kouse & Gary Klein. **"The tipester sumac text summarization evaluation"**. Published in proceeding of EACL, 1999.

Jum & Lin. "**summarization**". University of Maryland. published by Springer. 2009.

Julian Kupiek , Jan Pedersen & Francine. "**Trainable Document summarizer "**. In Research and Development in Information Retrieval, pages 68–73, 1995.

Kamal Sarkar. "**Using Domain Knowledge for Text Summarization in Medical Domain**". International Journal of Recent Trends in Engineering,Vol 1, No.1,May 2009.

Automatic Summarization System For Arabic Texts

Yusra Malik Dumamad *Suhad Muhajer Kareem*
Lecture

University of basrah / science collage

Yus_malik@yahoo. Com *suhad_althaher83@yahoo. Com*

Abstract

Automatic summarization plays an important role in light of the development of World Wide Web, which represents the most important ways to manage the large amount of text information that a user needs to read the text and by reducing the amount of text to help determine whether it has to do with the information they need or not. In this paper we proposed the automatic system for the Arabic text, trying to imitate human behavior so in writing the summary of the text provided. In this paper deals with method of a hybrid (statistical, linguistic), where we proposed several mathematical equations that merging a set of text properties to process the original text and get a find a summary coherent that contain useful information then was selected its efficient in the production of a summary is good and using the algorithm MMR has been removed redundant from sentences Summary produced. And then evaluated by comparing the fin Summary of the system with the summaries produced by experts in the Arabic language using ROUGE evaluation system by computing the precision, recall and F-Measure. The results were good and very encouraging , note that the systems summary of the Arabic language are very small compared with the rest of the languages. The system is programmed in Delphi 7 language.