التلخيص الاستخراجي للنصوص العربية

ريما السهلي

قسم الذكاء الصنعي، جامعة دمشق

ملخص

يوجد اليوم كمية هائلة من المستندات المتاحة عبر الإنترنت، لذا فإن معالجة النصوص للحصول على المعلومات بسرعة وكفاءة يصبح أمرًا صعبًا وهاماً جداً، يوضح هذا أهمية التلخيص التلقائي للنص، والتلخيص هو مهمة ضغط النص إلى نسخة أقصر، وتقليل حجمه مع الحفاظ في نفس الوقت على المعلومات الرئيسية ومعنى المقصود من النص. في هذه الورقة، نقدم شرح موجز لتقنيات تلخيص النص الاستخراجي، مع تطبيق بعض تقنيات التلخيص الاستخراجي على اللغة العربية.

الكلمات المفتاحية:

Automatic Document Summarization, Extractive Summarization Technique, Arabic Text Summarization.

I. المقدمة:

في عصر البيانات الضخمة، هناك انفجار في كمية البيانات النصية الصادرة من مصادر متنوعة، هذا الحجم من النص هو كنز لا يقدر بثمن للمعلومات لذلك لابد من تلخيص فعال لاستخراج المعلومات المفيدة، نظرًا لأن التلخيص اليدوي للنص يعد مكلفًا للوقت ومهمة شاقة بشكل عام، فإن أتمتة المهمة تكتسب شعبية متزايدة وبالتالي تشكل دافعًا قويًا للبحث الأكاديمي.

هناك تطبيقات مهمة لتلخيص النص في مختلف المهام المتعلقة بالبرمجة اللغوية العصبية مثل تصنيف النص، والإجابة على الأسئلة، وتلخيص النصوص القانونية، وتلخيص الأخبار، وتوليد العناوين، ويمكن دمج إنشاء الملخصات في هذه الأنظمة كمرحلة وسيطة تساعد على تقليل طول المستند.

التلخيص التلقائي للنص هو مهمة إنتاج ملخص موجز دون أي مساعدة بشرية مع الحفاظ على معنى المستند النصي الأصلي، إنه أمر صعب للغاية، لأننا عندما نلخص كبشر جزءًا من النص، فإننا نقرأه بالكامل عادةً لتطوير فهمنا، ثم نكتب ملخصًا يسلط الضوء على نقاطه الرئيسية، ونظرًا لأن أجهزة الحاسوب تفتقر إلى المعرفة البشرية والقدرة اللغوية، فإنها تجعل التلخيص التلقائي للنص مهمة صعبة للغاية.

تجري عملية التلخيص التلقائي للنص بواسطة تقنيتين أساسيتين هما: التلخيص الاستخراجي والتخليص التجريدي.

التلخيص الاستخراجي يلتقط الجمل مباشرة من المستند بناءً على أهميتها لتشكيل ملخص جديد، تعمل هذه الطريقة من خلال تحديد أقسام النص المهمة وتجميعها معًا لإنتاج نسخة مختصرة، وبالتالي فهي تعتمد فقط على استخراج الجمل من النص الأصلي.

ركزت معظم أبحاث التلخيص اليوم على التلخيص الاستخراجي، لأنه أسهل ويعطي ملخصات نحوية طبيعية ومقبولة.

وتهدف طرق التلخيص التجريدية إلى إنتاج ملخص عن طريق تفسير النص باستخدام تقنيات معالجة اللغة الطبيعية المتقدمة من أجل إنشاء نص أقصر جديد، قد لا تظهر أجزاء منه كجزء من المستند الأصلي، والذي ينقل المعلومات الأكثر أهمية من النص الأصلي، مما يتطلب إعادة صياغة الجمل ودمج المعلومات من النص الكامل لتوليد الملخصات مثل الملخصات المكتوبة بواسطة الإنسان عادة.

يوضح القسم 2 نوضح مسح موجز للأعمال السابقة وقسم 3 يوضح خطوات الحل المستخدمة، بينما يظهر القسم 4 التجارب التي أجريت، ثم في القسم 5 تظهر نتائج التجارب.

II. الأعمال السابقة:

A. أنواع تلخيص النصوص

يوضح Nenkova (2005) أن هناك نوعين من التلخيص[2]، يعتمد على نوع الدخل وهو تلخيص وثيقة واحدة وتلخيص متعدد الوثائق، في المقابل، اعتبر ALI GULIYEV (2009) أن أنواع التلخيص على النحو التالي:

1) استخلاص (استخراجي): حيث يتم اختيار بعض الجمل من النص الأصلي لتقديم نفس فكرة النص بشكل مختصر.

2)التجريدي: يمكن أن يؤدي الملخص إلى إنشاء جمل جديدة ليست جزءًا من النص الأصلي باستخدام معالجة اللغات الطبيعية العميقة المعالجة.

B. تلخيص النصوص العربية وتحدياتها

مع تزايد الأبحاث باللغة الإنجليزية تعانى باقي اللغات من نقص في العمل الجاد والمثمر ومنها اللغة العربية.

يعتبر إجراء تلخيص للنص العربي أمرًا صعبًا نظرًا لوجود العديد من التحديات مثل عدم وجود علامة التشكيل لمعظم المقالات العربية المكتوبة، والتي تتطلب تحليلًا متطورًا لمعرفة التشكيل الصحيح، مما يساعد في الحصول على المعنى الصحيح للكلمة والجملة، وسبب جعل اللغة العربية أكثر صعوبة من اللغات اللاتينية الأخرى التشكل المعقد، وقلة أدوات البرمجة اللغوية العصبية مقارنة باللغات الأخرى، وعدم وجود الحرف الأول الكبير الذي يشير إلى الأسماء، كما أن من تعقيد تحليل اللغة العربية أنه يحتوي على العديد من أشكال الحروف التي يمكن أن تصل إلى أربعة أشكال وأحيانا أكثر. بالإضافة إلى ذلك، يتم تغيير كتابة الحروف اعتمادًا على موضعها في الكلمة، فيما يتعلق بعملية الاشتقاق، هناك تعقيد يتطلب بذل جهد إضافي لإجراء عملية عالية الجودة لاستئصال الجذور[1].

C. الأبحاث السابقة في تلخيص النصوص العربية

تمت دراسة تقنيات وطرق التلخيص العربية الرائدة والتحقق من النتائج [1] ونذكر أبرز الأبحاث وأحدثها وترد على النحو التالي:

1. Haboush et al 2012، في دراسة التلخيص العربي باستخدام تقنية clustering technique, ، واستخدم الجذر بدلاً من الكلمة نفسها ، حيث يستخدمون نفس منهجية مراحل التلخيص باستثناء اعتماد مرحلة الترتيب rank على جذر الكلمات. تظهر النتيجة تحسنًا بنسبة 14٪ في الدقة و 9٪ في الاستدعاء ، بقيم 75.5٪ و 78.7٪ على التوالي.
2. Al-Abdallah, and Al-Taani(2017 استخدموا في بحثهم خوارزمية تحسين سرب الجسيمات e swarm optimization algorithm لتوليد تلخيص النص العربي اعتمادًا على نهج مستند واحد. تظهر دراسته استدعاء 54.44٪ ، دقة 58،82٪ ، ومقياس F 55.32٪.
3. (Waheeb et al. (2020 استخدم Clustering و Word2Vec لتلخيص النص العربي من مجموعة EASC وأظهرت النتائج 69.5٪ ، 60٪ ، 64.4٪
4. Al-Abdallah, and Al-Taani, (2019 طبق خوارزمية Firefly لاستخراج تلخيص النص العربي وأظهرت النتائج 60.14٪ و 57.32٪ و 57.52٪ للاسترجاع والدقة والقياس.
5. تم استخدام الخوارزمية الاستخراجية القائمة على الرسم البياني Graph-Based Extractive algorithm في دراسة Elbarougy, Behery, and KHATIB,(2020 ، كما قاموا بتطبيقه في تلخيص مستند واحد تظهر النتائج استدعاء 73.8٪ ، دقة 82.67٪ ، ومقياس F 76.37٪.

6. Elbarougy, Behery, and El Khatib,(2020 استخدام تلخيص النص العربي المستخرج باستخدام تعديل خوارزمية PageRank لإنشاء نسخة قصيرة من المستند الفردي المقدم من مجموعة EASC ، تظهر النتائج قيم 72.94٪ 68.75٪ 67.99٪ للاسترجاع و الدقة و القياس بأثر مستقبلي.
7. Abu Nada et al.(2020 يصفون استخدام AraBERT Model باستخدام تلخيص النص الاستخراجي نهج لتلخيص مستند واحد ، وتظهر النتائج 39٪ 90٪ 54٪ للاسترجاع والدقة و f قياس مستقبلي.
8. Jaradat, and Al-Taani, (2016 يستخدمان خوارزمية هجينة في وثيقة واحدة من مجموعة EASC ، وأظهرت النتائج 57.13٪ ، 56.58٪ ، 54.76٪ للاسترجاع والدقة و القياس بأثر مستقبلي

III. الطريقة Methodology:

تمر عملية تلخيص النصوص الاستخراجية بعدة مراحل تبدأ بالمعالجة المسبقة للنص، ثم تمثيل النص على شكل قيم رقمية واستخراج الميزات منه (feature extraction) وتليها مرحلة الاستفادة من هذه الميزات للحصول على ملخص المطلوب .

A. Preprocessing: تمت معالجة النص على عدة مراحل وهي:

1) تقسيم النص لجمل ثم إلى كلمات (tokenization) .
2) حذف كلمات التوقف(step word).
3) حذف علامات الترقيم والروابط.
4) عمل معالجة خاصة باللغة العربية (إزالة الحركات، إزالة التطويل، إزالة ما هو غير عربي، توحيد الهمزات).
5) تجذير الكلمات.

B. Feature extraction

وهي مرحلة تحويل الكلمات والجمل لقيم رقمية تعبر عنها نستطيع من خلالها التعامل معها، ويوجد عدة طرق لذلك منها ما هو إحصائي ومنها ما هو قائم على التعلم الآلي.

الطرق الإحصائي تعبر عن كل كلمة بقيمة رقمية تظهر ميزة مثل عدد مرات تكرار الكلمات في النص (تردد الكلمات، طول الجملة، موقع الجملة في النص، عدد القيم الرقمية التي تحتويها الجملة، تشابه بين الجمل).

أما الطرق القائمة على التعلم الآلي في تأخذ السياق بعين الاعتبار وتعطي الجملة قيمة مناسبة تعبر عنها، حيث يمكن تدريب نموذج تعلم آلي أو استخدام أحد النماذج المدربة مسبقاً.

C. Modeling extractive summary

توجد عدة طرق لاستخراج ملخص من الميزات المستخرجة ومن هذه الطرق ما هو قائم على الإحصاء مثل page rank Algorithm، Score-Based Algorithm، منها ما هو قائم على التعلم مثل K_mean Algorithm، نسعى إلى تجربة عدة خوارزميات وإيجاد أفضل نتائج.

1) Score-Based Algorithm

قد يعبر عنها بشكل Bog-of-word وTF-IDF لكنها جميعا قائمة على نفس المبدأ وهو حساب قيم رقمية تعبر عن الجمل ثم أخذ الجمل ذات القيم الأكبر لوضعها في الملخص [4].

2) Page rank Algorithm

تقوم هذه الخوارزمية على حسب التشابه بين كل جملتين من جمل النص ثم تشكل بيان graph تكون فيه الجمل هي عقد البيان والوصلات تحمل مقدار التشابه بين كل جملتين (عقدتين)، يتم ترتيب أهمية هذه الجمل بالاعتماد على أكثر جملة مشابهة لكامل جمل النص[5].

Algorithm 1. PageRank algorithm [6]

```
   Input: Weighted Graph G.
   Output: Scored Graph.
1  Configure N = Number of Nodes in G.
2  Current_Rank ← Double[N]
3  Temp_Rank ← Double[N]
4    Foreach n = 1 to N
5    Current_Rank[n] = 1/N
6      For i = 1: Number_of_Iterations
7    Foreach nd: G.Nodes
8      Temp_Rank[nd.index] = Calc_Page_Rank(nd)
9    Current_Rank = Temp_Rank
```

3) K_mean Algorithm

هي خوارزمية تعلم غير خاضعة لإشراف، يتم الاعتماد فيها المتوسط و وتقسم البيانات المررة لها إلى مجموعات جزئية Clustering، وفي تلخيص النصوص تقسم البيانات الجمل المررة إلى قسمين:
1. ينتمي للملخص 2. لا ينتمي للملخص.

المراجع:

[2]AlGhanem, Hani S., and Rashan H. Ajamiah. "Arabic text summarization approaches: A comparison study." *International Journal of Information Technology and Language Studies* 4.3 (2020).

[3]Abdelaleem, Nadeen M., HM Abdal Kader, and Rashed Salem. "A Brief Survey on Text Summarization Techniques." *IJ of Electronics and Information Engineering* 10.2 (2019): 103-116.

[4]Qaroush, Aziz, et al. "An efficient single document Arabic text summarization using a combination of statistical and semantic features." *Journal of King Saud University-Computer and Information Sciences* 33.6 (2021): 677-692.