

TEMPLATE BASED SUMMARIZATION FOR ARABIC DOCUMENTS

Prepared by

Ala'a Abdallah Al-Zoubi

Supervised by

Prof. Ghassan Kanaan

Abstract

Since the number of electronic documents increase quickly in the world, the need for faster techniques to assess of these documents arises. A summary is a brief representation of implicit text. To form an ideal summary, a full understanding of the document is primary. However, obtaining a full understanding is either impossible or difficult for computers. Therefore, selecting significant sentences from the original text and presenting these sentences as a summary setting the most common techniques in automated text summarization. This research focus on the NEWS article and what you want to do with the entire news article to enable the user to decide if they want to read the whole article or not, good summaries are rare, and too costly to produce manually, so the researcher writes a program to read a document collection and then produce a summary based on the designed template for this purpose

نماذج مرتكزة على تلخيص الوثائق العربية

إعداد

علاء عبدالله الزعبي

إشراف

الأستاذ الدكتور غسان كنعان

الملخص

إن عدد الوثائق الإلكترونية يزداد بسرعة في العالم، ولهذا تنشأ الحاجة إلى تقنيات لتقييم هذه الوثائق واسترجاعها بسرعة. ومن هذه التقنيات التلخيص وهو عبارة عن عرض موجز للنص الأصلي. ولإيجاد ملخص مثالي للوثائق، فإن الفهم الكامل للوثيقة أمر أساسي. ومع ذلك، فإن الحصول على فهم كامل للوثيقة أمر مستحيل أو صعب. لذلك، اختيار الجمل المهمة والأساسية من النص الأصلي وتقديم هذه الجمل كملخص يعدّ من التقنيات الأكثر شيوعاً في تلخيص الوثائق. يركز هذا البحث على المقالات الإخبارية، فبدلاً من قراءة كامل المقال الإخباري يستطيع المستخدمون أن يقرروا ما إذا كانوا يريدون قراءة المقال كاملاً أم لا، وذلك بقراءة الملخص القائم على النموذج المبني في هذه الرسالة. واحدة من أهم القضايا التي واجهتنا هي تلخيص المقالات المكتوبة باللغة العربية، واللغة العربية هي واحدة من أصعب اللغات في العالم وأكثرها تعقيداً، فهي تتكون من 28 حرفاً، تكتب من اليمين إلى اليسار إضافة إلى قواعدها النحوية واللغوية المتشعبة، وبالتالي فإن قليلاً من الأبحاث العلمية سلطت الضوء على استخراج النماذج الملخصة باللغة العربية. وبالتالي، فإن الفكرة الرئيسية من هذه الرسالة هي بناء ملخصات باللغة العربية قائمة على النماذج، وهذه العملية تمر في عدة مراحل وصولاً إلى النموذج النهائي المصنف حسب نوعه. ومن هذه المراحل، اختيار

المقالات الإخبارية وتدقيقها وادخالها على برامج استرجاع البيانات لاستخراج الخصائص الرئيسة من المقال الإخباري، مثل اسم الكاتب وتاريخ النشر ومكانه بعد ذلك تؤخذ هذه الخصائص لتوضع في نموذج مبني مسبقا في الرسالة لبناء تلخيص مفهوم المعنى وصحيح التركيب.