Automatic Arabic Text Summarization System Based on Semantic Features Extraction

Kathrein Abu Kwaik; Nabil M. Hewahi

Abstract

Recently, one of the problems arisen due to the amount of information and it's availability on the web, is the increased need for effective and powerful tool to automatically summarize text. For English and European languages an intensive works have been done with high performance and nowadays they look forward to multi-document and multi-language summarization. However, Arabic language still suffers from the little attentions and research done in this filed.

In our research we propose a model to automatically summarize Arabic text using text extraction. Various steps are involved in the approach: preprocessing text, extract set of feature from sentences, classify sentence based on scoring method, ranking sentences and finally generate an extract summary. The main difference between our proposed system and other Arabic summarization systems are the consideration of semantics, entity objects such as names and places, and similarity factors in our proposed system. The proposed system has been applied on news domain using a dataset obtained from Falesteen newspaper. Manual evaluation techniques are used to evaluate and test the system. The results obtained by the proposed method achieve 86.5% similarity between the system and human summarization. A comparative study between our proposed system and Sakhr Arabic online summarization system has been conducted. The results show that our proposed system outperforms the Shakr system.

Keywords: *Automatic Text Summarization, Feature Extraction, Manual Evaluation, Natural Language Processing.*

عنوان البحث:

النظام الإلكتروني في تلخيص المستندات العربيةاعتماداً على استخلاص الخصائص من حيث المعنى

ملخص:

في الآونة الأخيرة ظهرت العديد من المشاكل نتيجة ازدياد عدد المعلومات المنتشرة على الإنترنت، وأصبح الوصول إلى البيانات المطلوبة والمرادة من أصعب الأمور التي تواجه الباحثين.هذا أدى لظهور الحاجة إلى نظام محوسب آلي يقوم بتلخيص المستندات بشكل الكتروني وعرضها للمستخدم بحيث يستطيع تحديد ما إن كان المستند يفي بالغرض المطلوب أم لا.

إن العديد من أنظمة التلخيص الآلي تدعم اللغات الأوروبية وبالأخص اللغة الإنجليزية وتعطي نتيجة دقيقة جدا في انشاء ملخص عن اي مستند وقد تطور العمل على هذه اللغات لتصل إلى امكانية تلخيص عدة مستندات وبلغات مختلفة.إلا أنه وعلى الرغم من ذلك فإن هنالك قصور كبير في التعامل مع اللغة العربية في هذه المجالات ومازالت الأبحاث مستمرة في تطوير هذا المجال.

في هذا البحث سنقوم بعرض تصميم وانشاء لنظام تلخيص للمستندات العربية على اساس الاستخلاص من النص، هذا المقترح يتكون من عدة مراحل اهمها : مرحلة ماقبل المعالجة واستخراج الخصائص ومرحلة تصنيف الجمل النصية معتمدا على طريقة النقاط ومن ثم مرحلة ترتيب الجمل حسب أوزانها وقيمها وفي النهاية انشاء الملخص فيتم عرضه للمستخدم بطريقة سهلة وسريعة.

إن الإختلاف الرئيسي بين النظام المقترح وباقي الأنظمة العربية المستخدمة في التلخيص الآلي هو الاهتمام بعلم الألفاظ والمعاني سوياً واضافة نظام للتعرف على الاسماء والاماكن وبعض الاحداث الهامة من وجهة نظر الانسان وفحص مدى التقارب والتشابه مابين الجمل في المستند الواحد.

في عملية تقييم النظام وفحصه تم استخدام طريقة التقييم اليدوية وقد توصلنا إلى أن هناك تشابه مابين النظام وبين التلخيص البشري بنسبة 86.5 %. تم عمل دراسة مقارنة ما بين نظامنا وبين نظام صخر للتخليص الآلي. أظهرت النتيجة تفوق النظام المقترح عن نظام صخر.

الكلمات المفتاحية : تلخيص النصوص الآلي، استخلاص الخصائص، التقييم اليدوي، معالجة اللغة الطبيعية