نهج قائم على المحولات المدربة مسبقاً للتلخيص الاستخراجي والتجريدي للنص العربي

ياسمين عينيه

المشرفة : د.أمل المنصور

المستخلص :

التلخيص التلقائي للنص هو موضوع بحث بارز في معالجة اللغة الطبيعية بسبب تنوع وانتشار مصادر المعلومات على الإنترنت. من خلال هذه الدراسة درسنا نوعي التلخيص: الاستخراجي والتجريدي. تعتمد طريقة التلخيص الاستخراجي على اختيار أهم العبارات والجمل من نص الإدخال الرئيسي لإنشاء ملخص جديد دون إعادة تنسيق هذه العبارات والجمل. من ناحية أخرى ، يعتمد التلخيص التجريدي على تلخيص النص الأصلي بعبارات وجمل مختلفة تماماً. تم نشر العديد من الأعمال حول التلخيص التلقائي للغة الإنجليزية للعثور على المنهجيات الأكثر تقدماً والحصول على نتائج متقدمة. ومع ذلك ، فإن البحث في تلخيص النص العربي يتقدم ببطء أكثر بسبب طبيعة اللغة العربية والحاجة إلى المزيد من مجموعات البيانات المرجعية الأساسية. أظهرت العديد من نماذج اللغة المدربة مسبقاً مؤخرًا أداءً ممتازًا في العديد من مهام معالجة اللغة الطبيعية. لقد عملنا على ضبط ومقارنة أداء نموذج {AraBERT } الأساسي ونموذج {QARiB} ونموذج {AraELECTRA}. تم تدريب هذه النماذج باستخدام مجموعات البيانات العربية {KALIMAT} و {EASC} للتلخيص الاستخراجي للنص العربي. ثم تم تقييم الملخصات التي تم إنشاؤها باستخدام حزمة تقييم {ROUGE} باستخدام مقاييس {ROUGE-1} و {ROUGE-2} و {ROUGE-L}. تم تحقيق أفضل النتائج باستخدام نموذج {AraBERT} ، الذي حصل على {٠,٤٤} و {٠,٢٦} و {٠,٤٤} على مجموعة بيانات {KALIMAT}. بالإضافة إلى ذلك ، من أجل تلخيص النص التجريدي العربي ، استخدمنا محول تحويل النص إلى نص نموذج {T5} ، والذي أسفر عن نتائج جيدة. استخدمنا مجموعة بيانات من {٢٦٧٠٠٠} مقالة عربية لصقل {AraT5} ، النسخة العربية التي تم إطلاقها حديثاً. تم تقييم النموذج من خلال درجات {ROUGE-1} و {ROUGE-2} و {ROUGE-L} و {BLEU} ، وكانت النتائج {٠,٤٩٤} و {٠,٣٣٩} و {٠,٤٦٩} و {٠,٤٢٢٤} على التوالي. استخدمنا أيضاً مجموعة بيانات أخرى تحتوي على {٣٠٠٠٠٠ } مقالة وعناوين رئيسية وحققنا درجات التقييم التالية {٠,٥٣} و {٠,٣} و {٠,٣٦} و {٠,٤٨}. بالإضافة إلى ذلك ، كان نموذج {AraT5} متفوقاً على أحدث الأبحاث التي استخدمت نموذج التسلسل إلى التسلسل {Seq2Seq}.

Pre-trained Transformer-Based Approach for Extractive and Abstractive Summarization of Arabic Text

by

Yasmin Einieh

Advisor

Dr. Amal Almansour

Abstract:

Automatic Text Summarization (ATS) is a prominent research topic in Natural Language Processing (NLP) due to the variety and proliferation of information sources on the Internet. In this research study, we explored the ATS systems for two different approaches: extractive summarization and abstractive summarization. The extractive summarization method relies on selecting the most important phrases and sentences from the main entry text to create a summary without reformatting these phrases and sentences. Abstractive summarization, on the other hand, summarizes the original text in entirely newterms and sentences. There is plenty of research published on summarizing English text using more advanced methodologies to achieve advanced results. However, due to the nature of the Arabic language and the need for more basic reference datasets, research in Arabic text summarization is moving more slowly. Several pre-trained language models have recently shown excellent performance on many NLP tasks. For this reason, this study aims to experiment with different pre-trained models for summarizing the Arabic text. We finetuned and compared the performances of the base AraBERT model, the QARiB model, and the AraELECTRA model. These models were trained using the KALIMAT and EASC Arabic datasets for Arabic extractive text summarization. Then the generated summaries were evaluated with the ROUGE evaluation package using the ROUGE- 1, ROUGE-2, and ROUGE-L scales. The best results were achieved using the AraBERT model, which obtained 0.44, 0.26, and 0.44 on the KALIMAT dataset. In addition, for Arabic abstractive text summarization, we used the Text-to-Text Transfer Transformer (T5 model), which yielded good results. We used a dataset of 267,000 Arabic articles to finetune AraT5, the newly launched Arabic version. The model was evaluated through ROUGE-1, ROUGE-2, ROUGE-L, and BLEU scores, and the results were 0.494, 0.339, 0.469, and 0.4224, respectively. We also used another dataset containing 300,000 articles and headlines and achieved the following evaluation scores 0.53, 0.3, 0.36, and 0.48. In addition, the AraT5 model was superior to the most recent research using the Sequence-to-Sequence (Seq2Seq) model.