نظام محادثات نصي مبني على نموذج في الشبكات العصبية التسلسلية للهجة العربية الخليجية

اسم الطالب : تهاني فهد الشريف

المشرف : الدكتور معظم صديقي

الملخص

نظام المحادثات، أو نظام الحوار، هو نظام حاسوب لديه القدرة على الاستجابة للبشر تلقائيًا باستخدام اللغة الطبيعية. تقدم هذه الأنظمةاستجابات فورية ويمكنها في نفس الوقت مساعدة عدد غير محدود من المستخدمين. وجد تطوير أنظمة المحادثات باللغة العربية باهتمام أقل حتى الآن عند مقارنتها باللغات الأخرى بسبب تعقيد اللغة العربية،وجود عدة لهجات ونقص البيانات. وجدت الدراسات السابقة أن تطوير أنظمة المحادثات باللغة العربية يتم في الغالب باستخدام مطابقة الأنماط واسترجاع المعلومات، باستخدام مناهج التصنيف مع مصدر بيانات المجال المحدد. الدراسات في مجال تطوير أنظمة المحادثات مفتوحة المجال محدودة للغاية في مجال اللهجة العربية. استخدم هذا البحث بنية التعلم العميق، والمعروفة باسم الشبكة العصبية التسلسلية،لبناء نظام محادثات يستخدم اللهجة الخليجية العربية للرد. قام هذا البحث بصياغة مشكلة نظام المحادثات كمشكلة ترجمة آلية، وبالتالي،فقد تم بناء مصدر البيانات ليتناسب مع نموذج البيانات المستخدمة في تدريب مثل هذه الأنظمة من مجموعة التغريدات من موقع التواصل الاجتماعي تويتر. يختار هذا البحث لتقييم النصوص التي يكونها النظام احدى الطرق المستخدمة في تقييم نماذج الترجمة الآلية وباستخدام أشخاص مقيمين. بحسب الدراسات السابقة، فإن هذه الدراسة تعتبر الخطوة الأولى لتوليد النصوص من نماذج تعلم الألة "التسلسلية"

A seq2seq Neural Network based conversational Agent for Arabic Gulf Dialect

By Tahani Fahad Alshareef
Supervised by Dr. Muazzam Siddiqui

Abstract

A Conversational Agent (CA), or dialogue system, is a computer system that has the ability to respond to humans automatically using natural language. CAs offer instant responses and can concurrently assist a potentially unlimited number of users. The modeling of CAs in Arabic has so far received less attention when compared with other languages due to the complexity of the Arabic language, the existence of several dialects, and a lack of data resources. The literature contends that modeling a CA in Arabic mostly done using pattern-matching and information retrieval, employing classification approaches with a closed-domain data source. There is extremely limited research so far on modeling an open-domain CA in the Arabic dialect. This research has utilized a deep-learning architecture, known as the Seq2Seq neural network, to build a CA in the Arabic Gulf dialect. We formulated the CA problem as a machine translation problem and, therefore, built our corpus from tweets, in the post–reply format, to train and evaluate the model. We investigated the effects of pre-trained embeddings on the performance of the CA. For our evaluation, a Bilingual Evaluation Understudy (BLEU) score and human evaluators were used. The performance of the model was found to be comparable with existing deep learning models that have been trained on much larger corpora and in other languages. Our results present a promising first step towards building an open-domain CA in the Gulf Arabic dialect.