نظام أسئلة وإجابة آلي للبيانات النصية في اللغة العربية

توفيق زهير حسنين

المستخلص

تعتبر أنظمة إجابة الأسئلة QA من الأنظمة التي لها باع طويل في مجال الذكاء الاصطناعي، فهي تشمل الكثير من الضوابط والتي تمتد من النظريات الفلسفية وصولاً إلى نظريات قواعد البيانات، وقد تم التحقيق في عملية إجابة الأسئلة في هذا البحث من عدة جوانب بناءا على الاختلاف بين هذه الضوابط.

من ناحية أخرى يقدم هذا البحث نظاما مقترحا للبيانات النصية باللغة العربية، حيث تم تطوير نظام استوجاع المعلومات IR العربية، ومن المعلوم أن عملية البحث داخل جسم هائل من النصوص تعتبر مهمة صعبة وتستغرق وقتا طويلا للمستخدم العادي، لذا فإن تأسيس طريقة استرجاع للبيانات قد تكون فعالة جداً لكثير من المستخدمين.

في هذا البحث تم تطبيق النظام على حقل الحديث النب وي الشريف حيث افترضنا أن نظام الاسترجاع مقسم إلى مجموعة من المجلدات كل مجلد يحتوي على أحد كتب السنة التسعة، وكل كتاب يحتوي على مجموعة من الأبواب في حين أن كل باب يحتوي على مجموعة من الفصول الفرعية وهكذا دواليك.

يطبق هذا البحث نظم التصنيف النمطي بالاعتماد على التعلم الإحصائي من خلال موديول ماركوف الخفي (Hidden Markov Model) حيث تم بناء طراز واحد لكل باب في الكتاب وهي ما تسمى بعملية التدريب Training، لكن قبل القيام بهذا التدريب تم استخدام عملية هامة وهي عملية التجريد stemming حيث أن وظيفتها هي حذف المعلومات الصرفية وإعادة الكلمة إلى أصلها، ومن الجدير بالذكر أن اللغة العربية هي لغة ثرية بالمعاني حيث لها تصريف معقد جداً بعكس الكثير من اللغات الأخرى.

بعد عملية التجريد تم توليد قيمة مميزة feature vector لكل كلمة في النص حيث تم إنشاء قيمة للكلمات بالاعتماد على تكرارها داخل كل موضوع بواسطة تجميعهم إلى مجموعات لكل منها رقمها المميز، وقد استخدمنا خوارزمية k-means لأداء هذه الوظيفة.

بالرغم من استخدامنا للحديث النبوي الشريف إلا أن النظام يمكن أن يستخدم في أي مجال آخر، ولقد تم تطبيق عدد من التجارب بغية زيادة أداء النظام وقد تم تحقيق أكثر دقة تمكنا من إنجازها عند ٦٤% من صحة الإجابات.

Automatic Question Answering System for Arabic Language Textual Data

Tawfeq Z. Hasanain

Abstract

Question answering (QA) has a long tradition, involving many disciplines, ranging from philosophy to database theory. Depending on the discipline different aspects of the question answering process are investigated.

In this thesis, an Arabic Information Retrieval (IR) System has been implemented, we know that searching inside a large corpus is a hard and time-consuming task for the user, so that, establishing a way to retrieve the data to the user is very effective. The main concern is about the Prophetic Hadith. We have assumed that the corpus is divided into main topics and each one is divided into sub-topics and so on.

In another hand, this thesis presented the application of pattern recognition algorithm based on statistical learning, the Hidden Markov Model (HMM) which builds one model for each topic related trained texts and before training there is a processing step in any IR system which is stemming, which removes morphological information from the word. Stemming has a long tradition in document retrieval, and a variety of stemmers are available. The Arabic language is a highly inflected language and it has a complex morphology.

After stemming, and for training purpose a feature vector for each word in the corpus is generated. A new approach has been implemented, which creates the feature vector for the words from its frequency inside the topics, then labels are generated for the words by clustering them into groups and one label is given for all words in one cluster, the clustering process is used k-means algorithm witch classify or group our stems based on attribute/feature.

Although we used a Prophetic Hadith corpus, the system could be used in any other context, anyhow several experiments have been carried out in this research in order to increase the performance of our system and the highest possible accuracy accomplished in 64%.