Enhancing Arabic Named Entity Recognition Using Parallel Techniques

Prepared by:

Ziadoon Abdullah Otaiwi

Supervised by

Prof. Mohammed A. Otair

Abstract

Named entities recognition systems (Proper Names) are used in the development of many other natural language processing applications such as information retrieval, questions answering, machine translation, and others; because it provides important information about the text and is used as discriminatory features to improve the performance of these applications in carrying out their tasks.

There are many research works that have built techniques to identify named entities in most languages spoken around the world. Despite this, there is a paucity of published research in the field of identifying the named entities from texts written in Arabic. This is due to the fact that the Arabic language has a specificity regarding the complexity of spelling and morphology, which is an obstacle to the development of a technique to identify the names of the Arabic entities or the so-called Arabic Named Entities Recognition system (ANER).

This thesis presented the experiments conducted to identify the appropriate technique to design a robust and reliable system for identifying Arabic entities. The appropriate technique will have the ability to recognize named entities within Arabic texts different sizes. For this purpose, this study focuses on the most common state-of-art in the field of identification of Arabic named entities, then a comparison was made between five of the most famous tools that interested in identifying the Arab entities, after that, an integration process for each of two tools together was applied to get 10 different parallel techniques. The results of the comparison between the tools showed that Rosette achieved the best results followed by Madamira, while the worst performance results were in the gate tool and for parallel systems, the R-F (combining Rosette and Farasa) achieved the best performance with an accuracy better than the individual tools. Also, when comparing the values of the Recall scale, where some parallel systems achieved a great advantage compared to the best results achieved on the level of each tool alone, the results were about 0.985%, 0.979%, and 0.975 for short, Medium and long texts respectively.

تحسين التعرف على الكيانات الاسمية العربية باستخدام تقنيات متوازية

إعداد

زيدون عبد الله عطيوي

المشرف

الأستاذ الدكتور محمد عطير

الملخص

تستخدم انظمة التعرف على الكيانات الاسمية (اسماء الاعلام) في تطوير العديد من تطبيقات نماذج معالجة اللغات الطبيعية الأخرى مثل استرجاع المعلومات، واجابة الاستفسارات، والترجمة الآلية، وغيرها. وذلك لأنها تقدم معلومات هامة عن النص والتي يتم استخدامها كميزات دلالية لتحسين اداء هذه التطبيقات عند اجراء مهامها.

هناك العديد من الاعمال البحثية التي قامت ببناء نماذج للتعرف على اسماء الاعلام في معظم اللغات المستخدمة في مختلف انحاء العالم. وبالرغم من ذلك، فهناك ندرة في عدد الابحاث المنشورة في مجال التعرف على اسماء الاعلام من النصوص المكتوبة باللغة العربية وذلك بسبب ان اللغة العربية لديها خصوصية فيما يتعلق بتعقيد الإملاء والمورفولوجيا، الامر الذي يشكل عقبة امام تطوير نماذج للتعرف على اسماء الكيانات العربية او ما يعرف باسم:

"Arabic Named Entities Recognition system (ANER)"

هذه الرسالة تقدم العديد من التجارب التي أجريت من أجل تحديد النموذج المناسب لبناء نظام قوي وموثوق به للتعرف على اسماء الاعلام العربية. وسيكون لهذا النظام القدرة على تحديد اسماء الكيانات وتصنيفها ضمن نصوص عربية متفاوتة في احجامها. لهذا الغرض، وتم دراسة أحدث ما توصلت إليه التكنولوجيا في مجال تحديد الأسماء العربية. ثم أجريت مقارنة بين خمسة من أكثر الأدوات شهرة واهتماما بالتعرف على اسماء الاعلام العربية، وبعد ذلك قمنا بعملية دمج على مستوى كل اداتين معا للحصول على 10 تقنيات متوازية مختلفة.

وأثبتت نتائج المقارنة بين الأدوات ان Rosette حققت افضل نتائج تليها Madamira بينما كانت أسوء نتائج عند الاداة Gate . وبالنسبة للأنظمة المتوازية فقد حقق R–F (الذي يجمع بين Rosette وFarasa) افضل أداء بنسبة تحسين في الدقة أذا ما قورنت بنتائج كل أداة على حدة. ايضاً عند مقارنة قيم مقياس الRecall ، حيث حققت بعض الأنظمة المتوازية تفوق كبير مقارنة بأفضل نتيجة ظهرت على مستوى كل أداة لوحدها، فكانت النتائج حوالي 0.985% ,0.979% و0.975% للنصوص القصيرة والمتوسطة والطويلة على التوالي.