

Republic of Yemen
Ministry of Higher Education and Scientific Research
University of Science and Technology
Deanship of Postgraduate Studies and Scientific Research



A Framework for Classical Arabic Parsing Using Deep Learning

A Thesis Submitted by
Wadee Abdualhafiz Abdualaziz Nashir

To

**The Council of Postgraduate Studies and Scientific Research,
in Partial Fulfillment of the Requirements for the Degree of Doctor of
Philosophy in Computing**

Supervisor

Dr. Abdulqader M. A. Mohsen
Associate Professor of Computer Science
University of Science and Technology

July, 2020AD

Sana'a

Dhul Qi'dah, 1441AH



A Framework for Classical Arabic Parsing Using Deep Learning

Abstract

Arabic language is one of the most challenging languages for parsing because of the rich morphology and free word order system and the depth of the syntactical structure. This challenge increases with Classical Arabic, which is the old form of the Arabic language with more complex syntax. Moreover, Classical Arabic suffers from a lack of resources for parsing such as Treebank. Accordingly, this thesis has to main objectives 1) to the extending of the only Treebank of Classical Arabic; 2) and to develop a new framework for parsing. To achieve the first objective, the three layers within Quranic Treebank have been expanded. Various new algorithms have been introduced to achieve layers expansion. In the orthographical layer, a new script has been added that fits all the Classical Arabic texts instead of the existing script. Instead of the current numbering based on the concept of division of the Quranic Treebank, a new sequential numbering of sentences was added. In the morphological layer, the annotation scheme was moved towards more fine-grained tags by adding new tags and dividing some existing tags. Besides, approximately 0.73% of the morphological annotation was corrected, thus moving to a higher level in the gold-standard morphological annotation. In the syntactical layer, the entire layer was built according to a hybrid representation of the phrase-structure and dependency relations. To formalize this hybrid representation, the CoNLL-U format has been expanded. To achieve the second goal, this thesis presented a deep model for Classical Arabic parsing contains carefully selected standard techniques based on the best existing literature. The pipeline model is organized into three lines. In the pre-processing line, word embedding is pre-trained to achieve dense features representation. Also, the elliptical and phrase construction has been pre-processed to achieve a pure representation of sentences, the purpose of this is to make training data standard for dependency parsing. In this regard, the thesis provides new resources that have been introduced to Classical Arabic, the largest corpora for NLP tasks, and two models (Skip-gram and CWOB) for word embedding.

Republic of Yemen
Ministry of Higher Education and Scientific Research
University of Science and Technology
Deanship of Postgraduate Studies and Scientific Research



In deep parsing lines, simple and effective parser features are extracted by BiLSTM training along with the parser algorithm. Syntactic units are represented in sequence with their contexts by using Deep-BiLSTM vectors. Propagating errors are reduced through Oracle Dynamic and an aggressive-exploration scheme. We introduce a new transition system that allows direct syntax information to be used in transition decisions. In the pre-processing line, a deep visual representation was presented that reflected the complexities of the Classical Arabic structure. To remove the conflicts within the grammatical structure of the sentence, a semantic dimension describing syntax has been added. Finally, we demonstrate the effectiveness of the pipeline parsing model by comparing it with previous work. The proposed framework has proven to be the best in all different performance measures.

Candidate

Wadee Abdualhafiz Abdualaziz Nashir

Supervisor

Dr. Abdulqader M.A. Mohsen
Associate Professor of Computer Science
University of Science and Technology

July, 2020AD

Dhul Qi'dah, 1442AH



الجمهورية اليمنية
وزارة التعليم العالي والبحث العلمي
جامعة العلوم والتكنولوجيا
عمادة الدراسات العليا والبحث العلمي

في خط التحليل العميق ، يتم استخراج ميزات المحلل اللغوي البسيطة والفعالة من خلال تدريب BiLSTM جنبًا إلى جنب مع خوارزمية المحلل اللغوي. يتم تمثيل الوحدات النحوية بالتسلسل مع سياقاتها باستخدام متجهات Deep-BiLSTM. يتم تقليل أخطاء الانتشار من خلال مكون Oracle Dynamic ومخطط الاستكشاف (aggressive-exploration). تم تقديم نظام انتقال جديدًا يسمح باستخدام معلومات بناء الجملة في قرارات الانتقال بشكل مباشر. في خط المعالجة اللاحقة ، تم تقديم تمثيل مرئي عميق يعكس تعقيدات هياكل العربية الفصحى. ولإزالة التعارضات داخل البنية النحوية للجملة. كذلك، تمت إضافة بُعد دلالي يصف التركيبات النحوية. أخيرًا ، نوضح فعالية نموذج تحليل خط الأنابيب من خلال مقارنته بالأعمال السابقة. لقد أثبت إطار العمل المقترح أنه الأفضل في جميع مقاييس الأداء المختلفة.

المشرف

أ.د. عبدالقادر محمد عبدالرحمن محسن

أستاذ علوم الحاسوب المشارك
جامعة العلوم والتكنولوجيا

الطالب

وديع عبد الحافظ عبد العزيز ناشر



إطار عمل لتحليل اللغة العربية الفصحى باستخدام التعلم العميق

المُلخص

اللغة العربية هي واحدة من أكثر اللغات تحدياً للتحليل بسبب التشكل الغني ونظام ترتيب الكلمات الحر وعمق البنية النحوية. يزداد هذا التحدي مع اللغة العربية الفصحى ، وهي الشكل القديم للغة العربية مع تركيب أكثر تعقيداً. علاوة على ذلك ، تعاني اللغة العربية الفصحى من نقص في الموارد اللازمة للتحليل مثل البنوك اللغوية. وبناءً على ذلك ، فإن هذه الأطروحة تهدف إلى تحقيق الهدفين الرئيسيين التاليين (1) توسيع الشجرة الوحيد للغة العربية الفصحى. (2) وضع إطار عمل جديد للتحليل اللغوي. ولتحقيق الهدف الأول تم توسيع الطبقات الثلاث داخل الشجرة القرآنية. وتم إدخال خوارزميات جديدة مختلفة لتحقيق التوسيع للطبقات. في الطبقة الإملائية ، تمت إضافة نص جديد يتناسب مع جميع النصوص العربية الفصحى بدلاً من النص الحالي. كذلك، بدلاً من الترقيم الحالي القائم على مفهوم تقسيم الشجرة القرآني ، تمت إضافة ترقيم تسلسلي جديد للجملة. في الطبقة المورفولوجية ، تم نقل مخطط التعليقات التوضيحية نحو المزيد من العلامات الدقيقة عن طريق إضافة علامات جديدة وتقسيم بعض العلامات الموجودة. إلى جانب ذلك ، تم تصحيح ما يقرب من 0.73% من التعليق التوضيحي الصرفي ، وبالتالي تم الانتقال إلى مستوى أعلى في التعليق التوضيحي الصرفية القياسية. في الطبقة النحوية ، تم بناء الطبقة بأكملها وفقاً لتمثيل هجين شمل بنية العبارة وعلاقات التبعية. لإضفاء الطابع الرسمي على هذا التمثيل الهجين ، تم توسيع تنسيق CoNLL-U. بالنسبة لتحقيق الهدف الثاني ، قدمت هذه الأطروحة نموذجاً عميقاً لتحليل اللغة العربية الفصحى يحتوي على تقنيات قياسية مختارة بعناية اعتمدت على أفضل الأدبيات الموجودة. تم تنظيم نموذج خط الأنابيب في ثلاثة خطوط. في خط المعالجة المسبقة ، تم تدريب تضمين الكلمات بشكل مسبق لتحقيق تمثيل ميزات كثيفة. أيضاً ، تمت معالجة البناء التقديري وهياكل العبارات بشكل مسبق لتحقيق تمثيل اعتماد نقي للجملة ، والغرض من ذلك هو جعل بيانات التدريب معياراً لتحليل التبعية. في هذا الصدد ، توفر الأطروحة موارد جديدة تم تقديمها إلى اللغة العربية الفصحى ، وهي أكبر مجموعة لمهام البرمجة اللغوية العصبية ، ونموذجين (Skip-gram و CWOB) لتضمين الكلمات.



الجمهورية اليمنية
وزارة التعليم العالي والبحث العلمي
جامعة العلوم والتكنولوجيا
عمادة الدراسات العليا والبحث العلمي

إطار عمل لتحليل اللغة العربية الفصحى باستخدام التعلم العميق

رسالة مقدمة من الطالب
وديع عبد الحافظ عبد العزيز ناشر

إلى

مجلس الدراسات العليا والبحث العلمي، وهي جزء من متطلبات الحصول على درجة دكتوراه
الفلسفة في الحوسبة

المشرف
أ.د. عبدالقادر محمد عبدالرحمن محسن
أستاذ علوم الحاسوب المشارك
جامعة العلوم والتكنولوجيا-اليمن