University of Mosul
Collage of Computer Sciences
and Mathematics

Transforming Unstructured Text to Semantic Representations Based on Ontology

A Thesis Submitted By

Mustafa Nabeel Salim Abdulhadi

M.SC. / Thesis

Computer Science

Supervised By

Dr. Ban Shareef Mustafa

2021 A.D. **1443 A.H.**

Abstract

With the development of the Internet and the emergence of new technologies, data grows rapidly in a huge number in different shapes like news, articles, health care, social media ,etc. Most of this data comes in unstructured forms like natural speech and text that applications and users cannot get a benefit from it. Information Extraction (IE) which seeks to extract information that can be processed by computer from natural language text becomes a significant field and it took attention by researchers in the field of Data Science, Artificial Intelligence and others. Ontology Based Information Extraction (OBIE) is a partial field of IE which extracts data based on an ontology that presents specific concepts and relationships in a particular domain. In this thesis, model prototype Unstructured Text to Knowledge Base (UTtoKB) has been built which extracts semantic relationships from an unstructured text based on ontology. This model is a pipeline steps based on natural language processing (NLP) tasks and tools like Coreference Resolution, Named Entity Recognition, Semantic Role Labeling and Part of Speech Tagging. WordNet has been used as a tool to measure similarities between entities in order to convert them into ontology concepts and properties and populate them. It gives results of 75%, 70% and 72.41% for precision, recall and F1-score respectively.

جامعــة الموصــل
كليــة علوم الحاسوب
والرياضيات

تحويـــل النص غير المهيكل الى تمثيــل دلالـي بالاعتماد على علم الانطولوجيا

مصطفى نبيل سالم عبدالهادي

رسالة ماجستير

علـوم الحاسـوب

بإشراف

د. بان شريف مصطفى

١٤٤٣ هـ
٢٠٢١ م

الخلاصة

مع تطور الإنترنت وظهور تقنيات جديدة ، اصبح نمو البيانات بسرعة و بأعداد هائلة و بأشكال مختلفة مثل الأخبار والمقالات والرعاية الصحية ووسائل التواصل الاجتماعي إلخ ... تأتي معظم هذه البيانات في أشكال غير منظمة لا يمكن للتطبيقات والمستخدمين الاستفادة منها. لذلك أصبح حقل Information Extraction ،والذي يسعى لاستخراج المعلومات التي يمكن معالجتها بواسطة الكمبيوتر من نصوص اللغة الطبيعية، مجالًا مهمًا وقد حظي باهتمام الباحثين في مجال علوم البيانات والذكاء الاصطناعي وغيرهم. يعتبر Ontology Based Information Extraction مجال جزئي من Information Extraction ويعمد الى استخراج البيانات بناءً على انطولوجي والذي يقدم مفاهيم وعلاقات محددة في مجال معين.

في هذه الرسالة ، تم بناء نموذج أولي يسمى (UTtoKB) يستخرج العلاقات الدلالية من نص غير منظم قائم على انطولوجي. هذا النموذج عبارة عن سلسلة من الخطوات المعتمدة على مهام وأدوات معالجة اللغة الطبيعية Natural Language Processing مثل Named Entity Recognition و Coreference Resolution و Semantic Role Labeling و Part of Speech tagging. لقد تم استخدام WordNet كأداة لقياس أوجه التشابه بين الكيانات من أجل تحويلها إلى مفاهيم وخصائص الأنطولوجي وتعبئتها. اعطت النموذج نتائج ٧٥% و ٧٠% و ٧٢,٤١ بالتعاقب لكل precision و recall و F1-score.

.