International Journal of Applied and Natural Sciences (IJANS)
ISSN (P): 2319-4014; ISSN (E): 2319-4022
Vol. 7, Issue 2, Feb – Mar 2018; 65-86

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

A PROPOSED MODEL FOR EXTRACTING INFORMATION FROM ARABIC-BASED CONTROLLED TEXT DOMAINS, DISCUSSING THE INITIAL MODEL STEPS

Mohammad Fasha, Nadim Obeid & Bassam Hammo

Research Scholar, Department of Computer Science, King Abdulla II School for Information Technology, The University of Jordan, Amman

ABSTRACT

Information extraction from Arabic as well as other languages text is commonly implemented over restricted text domains. Approaching open text domains is challenging, because of the syntactic, semantic and pragmatics ambiguities and variations in text. For the purpose of approaching more relaxed versions of Arabic text domains, Fasha et al. (Fasha et al. 2017) presented a high-level description fora proposed work methodology that can establish a model for extracting information from controlled text domains. In that work, controlled text domains were defined as the text domains that are not restricted in their linguistic features or their knowledge types yet they are not very unanticipated in these respects. In this paper, we discuss that work methodology and its implementation in more detail. Our discussion includes the initial phases of the methodology which covers the corpus preparation processes including its selection, analysis and annotation using a custom morpho-syntactic Part-of-Speech tagging scheme, we also discuss the designing of the supporting knowledge-base model which will be used to represent and process the extracted information. The information extraction algorithm itself shall be presented in a future work.

***KEYWORDS**: Narabic Natural Language Processing POS Tagging Ontology Based Information Extraction Description Logic*

Article History

***Received:** 13 Jan 2018 | **Revised:** 11 Dec 2017 | **Accepted:** 03 Feb 2018*