

Abstract

Since the Quran is the primary holy book for Muslims, information extraction research on Quranic texts, especially in a form of People Entity Extraction, is an important task for further Quran and Tafseer understanding. The challenges in extracting people entities from the Quranic text is that many verses have a complex structure, such as nested entities, making it crucial to build a system that can extract the entity automatically, accurately, and quickly. People Entity Extraction on Quran itself is a task that aims to extract people entities in a sentence or verse, such as the name of a person, the name of a group, etc. on the Quranic texts. Example of input taken from snippet Surah Al-Baqarah verse 46 which reads “Those who believe that they will meet their Lord and that they will return to him” from that input the people entity extraction system is expected can identify people entities i.e. “Those who believe that they will meet their Lord”. Currently, People Entity Extraction research for the Quran has not been widely carried out, only a few algorithms with scattered results have been conducted. In this research, we then use several supervised models which are Conditional Random Field (CRF), BiLSTM-CRF, and a pre-trained deep learning model based on IndoBERT transformers. We apply and perform a comparative analysis for the performance of those several models. We found out that deep learning based model, namely BiLSTM-CRF perform best at extracting people entities, whilst probabilistic based model, namely CRF, had difficulty in extracting people entities, specifically nested people entities.

Keywords: Quran, People Entity Extraction, Supervised Learning, Comparative Analysis, Nested Entities