Abstract

The development of science is in line with the increasing number of publications in the scientific literature. Each piece of scientific literature contains information that is important to study in order to support the progress of research. Scientific literature, however, has some differences when being compared to other written work, so the processing of information in scientific literature also requires special techniques. Therefore, it is important to develop methods and automate the processing of scientific literature. This research aims to implement entity extraction in Indonesian scientific articles with Hidden Markov Model (HMM) and Conditional Random Fields (CRF) methods and analyze the results. The extracted entities consist of dataset, task, method, and evaluation metric. In this research, a dataset consisting of articles on the topic of informatics has also been built. Based on evaluation results, the CRF model outperforms the HMM model, with the F1-score of the CRF model being 0.71 and the F1-score of the HMM model being 0.65.

Keywords: CRF, entity extraction, HMM, information extraction, scientific literature