ABSTRACT

The branch of computer science that studies how computers can understand and analyze human language is the branch of Natural Language Processing (NLP). As a branch knowledge that understands the meaning of sentences given from language, computers have various tasks that can be done. Therefore, there are also many topics in the field of NLP that divide these tasks so that it is easy to reach solutions for the computer to carry out their tasks. One of the common topics in NLP that will also be discussed in this final project is Named Entity Recognition (NER).

The NER task is already common and has been widely researched, especially in the English language. However, there is a task that is part of NER which is still not as common as NER itself to be studied, that is the Nested Named Entity Recognition (Nested NER). There is a short difference between them, that is the recognition of entities in a sentence can be nested. For example, Jalan Ir. Soekarno is not only a location entity but also a nested entity in it, which is the word Ir. Soekarno as a person entity. There is one method that is most often used in several studies that have been carried out for Nested NER, this method is the span-based method. However, because some of its drawbacks such as computation and accuracy in forming spans from words, there is one method that was discovered in 2021, namely the Sequence-To-Set Network method.

Based on the results of the research for this method, this method beats accuracy in performance by 0.50% - 2.99% against the span-based method based on different datasets. This can be achieved with a similar concept to seq2seq which uses an encoder and decoder layer but with different layers and outputs. The encoder will encrypt the input sentence with a concatenation of variety different embeddings. Then the results will be passed to a decoder layer which has knowledge of self and cross-attention. Part of this decoder takes inspiration from the architectural form of Transformers. So, the output of the decoder can be a set containing left and right word boundaries and also the predicted entity type. Another thing that supports Sequence-To-Set Network to be an efficient method is the selection of a loss function based on bipartite matching with the Hungarian algorithm. The goal of this final project is to produce a good method for identifying an entity named/NER in Indonesian with a politically-dominated CNN Indonesia news dataset.