ABSTRAK

Cabang ilmu komputer yang mempelajari bagaimana caranya computer dapat memahami dan menganalisa bahasa manusia adalah cabang ilmu Pengolahan Bahasa Alami atau *Natural Language Processing* (NLP). Sebagai ilmu yang memahami arti dari kalimat yang diberikan dari bahasa, komputer memiliki beragam *task* yang bisa dilakukan. Karena itu ada banyak topik juga dalam bidang NLP yang membagi task-task tersebut agar mudah untuk mencapai solusi-solusi komputer melaksanakan tugas mereka. Salah satu topik dalam bidang NLP yang umum dan juga akan dibahas di tugas akhir ini adalah Named Entity Recognition (NER).

Task NER memang sudah umum dan banyak diteliti, terutama di Bahasa Inggris. Namun ada task merupakan bagian dari NER yang masih belum seumum NER sendiri untuk diteliti yaitu *Nested* Named Entity Recognition. Perbedaan yang cukup singkat yaitu pengenalan entitas dalam suatu kalimat bisa bersarang. Contohnya Jalan Ir. Soekarno bukan saja entitas lokasi namun juga terdapat entitas bersarang didalamnya yaitu Ir. Soekarno sebagai entitas orang. Terdapat satu metode yang paling sering digunakan dalam beberapa penelitian yang sudah dilakukan untuk Nested NER, metode ini metode *span-based*. Namun karena beberapa kekurangannya seperti komputasi dan akurasi dalam membentuk span memberi suatu halangan, ada satu metode yang baru ditemukan pada tahun 2021 yaitu metode Sequence-To-Set Network.

Berdasarkan hasil penelitiannya metode tersebut, metode ini mengalahkan akurasi dalam performa sebanyak 0,50% - 2,99% terhadap metode span-based dengan dataset berbeda-beda. Hal ini dapat dicapai dengan konsep yang mirip dengan *seq2seq* yaitu menggunakan *encoder* dan *decoder layer* namun dengan isi dan output berbeda. Encoder akan melakukan enkripsi terhadap kalimat input menjadi berbagai macam embedding yang berbeda-beda. Kemudian hasil itu akan dilewatkan decoder layer yang memiliki ilmu dari *self* dan *cross-attention*. Bagian dari decoder ini mengambil inspirasi dengan bentuk arsitektur Tranformers. Sehingga output dari decoder bisa berupa sebuah set yang berisi batasan kata kiri dan kanan dan juga jenis entitas yang diprediksikan. Hal lain yang mendukung Sequence-To-Set Network untuk menjadi metode yang efisien adalah pemilihan *loss function* berdasarkan *bipartite matching* dengan algoritma Hungarian. Tugas akhir ini memiliki tujuan untuk menghasilkan metode yang baik untuk pengenalan entitas bernama/NER dalam bahasa Indonesia dengan dataset berita CNN Indonesia berdomai politik.