ABSTRAK

Part-of-speech (PoS) *tagger* merupakan salah satu *task* dalam bidang *natural language processing* (NLP) sebagai proses penandaan kategori kata (*part-of-speech*) untuk setiap kata pada teks kalimat masukan. *Hidden markov model* (HMM) merupakan algoritma *PoS tagger* berbasis probabilistik, sehingga sangat tergantung pada *train corpus*. Terbatasnya komponen dalam *train corpus* dan luasnya kata dalam bahasa Indonesia menimbulkan masalah yang disebut *out-of-vocabulary* (OOV) *words*. Untuk mengatasi permasalahan tersebut dibutuhkan sebuah metode yaitu Analisis Morofologi. Penelitian ini membuat dua sistem yaitu *PoS tagger* HMM menggunakan metode Analsis Morfologi (AM) dan *PoS tagger* HMM tanpa AM, dengan menggunakan *train corpus* dan *testing corpus* yang sama. *Testing corpus* mengandung 30% tingkat OOV dari 6.676 token atau 740 kalimat masukan. Hasil yang diperoleh dari sistem HMM saja memiliki akurasi 97.54%, sedangkan sistem HMM dengan metode analisis morfologi memiliki akurasi tertinggi 99.14%.

Kata kunci: *bahasa Indonesia, natural language processing, part-of-speech tagging, hidden markov model, morphological analysis, out-of-vocabulary.*

ABSTRACT

Part-of-speech (PoS) tagger is one of tasks in the field of natural language processing (NLP) as the process of part-of-speech tagging for each word in the inputed sentence. Hidden markov model (HMM) is a probabilistic based PoS tagger algorithm, so it really depends on the train corpus. The limited components in the train corpus and the breadth of words in the Indonesian language pose a problem called out-of-vocabulary (OOV) words. To overcome this problem, a method is needed, namely Morophological Analysis. This research includes developing two systems, those are PoS tagger HMM using Morphological Analysis (AM) method and HMM PoS tagger without AM, using the same train and testing corpus. Testing corpus contains 30% OOV level out of 6,676 tokens or 740 sentences. The result obtained from the HMM system has 97.54% of accuracy, while the HMM system with morphological analysis method has 99.14% as it's highest accuracy.

Keywords**:** *bahasa Indonesia, natural language processing, part-of-speech tagging, hidden markov model, morphological analysis, out-of-vocabulary.*