ABSTRAK

Bahasa Indonesia merupakan bahasa resmi nasional, identitas bangsa dan lambang kebanggaan nasional. Namun, karena ketidaksengajaan maupun ketidaktahuan akan kaidah Bahasa Indonesia yang benar, sering terjadi penyimpangan dalam penulisan maupun pengucapan, salah satunya adalah struktur kalimat yang rancu. Untuk itulah dilakukan penelitian mengenai deteksi kalimat bahasa Indonesia berstruktur rancu, menggunakan metode *natural language processing* (NLP), *labeled sequential pattern* (LSP) dan klasifikasi. Dikumpulkan data kalimat rancu dan data kalimat tidak rancu dengan rasio 1:1. Semua data tersebut diproses dengan *part-of-speech tagger* hingga menghasilkan sekuen *tag* untuk setiap kalimat. Kemudian dilakukan ekstraksi LSP dari sekuen-sekuen *tag* tersebut mengunakan *sequential pattern mining.* LSP yang didapat dari setiap *instance* data digunakan sebagai *feature* masukan pada model *classifier.* Karena menggunakan *part-of-speech* sebagai komponen utamanya, sistem yang dihasilkan sangat tergantung pada *train corpus* yang digunakan pada proses *part-of-speech* tersebut. Setelah melakukan beberapa modifikasi pada *train corpus,* sistem memiliki akurasi sebesar 65.60% dan masih memerlukan pengembangan lebih lanjut, misalnya dengan menambahkan *feature* NLP lain seperti probabilitas *parsing*, mengkombinasikan LSP dengan *rule-based,* atau menambahkan *feature language model*.

Kata kunci: *part-of-speech tagging, bahasa Indonesia, sequential pattern, classification, feature extraction, grammatical error.*

ABSTRACT

Bahasa is a formal language in Indonesia, as well as the people's identity and a pride symbol. However, there are still many people using Bahasa, make some mistakes both in writing and speaking, one of it are grammatical error. Probably it because they are unintentionally did the mistakes or indeed they do not know what is the right one. That, is one of some more reasons to do the research about grammatical error detection in Bahasa on the structural/syntaxtycal point of view. Natural language processing (NLP), labeled sequential pattern (LSP) and classification are used in this research. The methods of research is in the following order: 1) collect Bahasa sentences, whether its grammatically error or grammatically correct, until the ratio is 1:1; 2) all of the collected sentences are proceed by part-of-speech tagger, that way the sentences are become sequence of tags; 3) then, sequential pattern mining extracted LSP from the sequence-of-tags database, and the LSP from each instances of the data are transformed into feature for classification input. The system use part-of-speech as the primary component. It turn out that the train corpus for the part-of-speech has the great effect in this methods. After modified the used train corpus several times, the system got 65.60% accuracy and the continuation development is still needed.

Keywords: *part-of-speech tagging, bahasa Indonesia, sequential pattern, classification, feature extraction, grammatical error.*