1. Introduction

Spelling errors occur in many kind of Indonesian texts regardless of their formality (whether they are formal texts or informal texts). Spelling errors are very common in human-generated texts and the cause of them can vary from accidental hand or mind slips to writer's lack of spelling knowledge. For reading purpose, spell detecting and correcting might be trivial for humans since humans can naturally detect patterns easily that make them able to read and understand texts even though the texts have spelling errors, while machines cannot do it without being instructed or trained to.

In Natural Language Processing (NLP) applications, data normalization is very important [1] because it can improve the performance and the accuracy of NLP applications. By performing spell checking before other normalization task for various NLP applications such as information retrieval, machine translation, text classification, and opinion mining, spell checking can reduce out-of-vocabulary (OOV), reduce the size of bag of words representation, and produce better stemming or lemmatization result. For humans, spell checking can help when they are writing texts that must contain no mistakes (often texts in formal context) or for a better readability of texts.

Spell checking is a challenging task because there are many error possibilities. The basic approach or detecting spelling errors, is using dictionary lookup. The problem arise when there are words that are not spelling errors but detected as errors because we could not find them in the dictionary or when there are named entities that are detected as spelling errors. Another approach for correcting spelling errors is rule-based approach and the challenge lies on many error possibilities, therefore we have to define many set of rules.

There are several spell checkers for Indonesian text that uses different approaches. The first study, proposed a rule-based approach that uses similarity measure and forward reversed dictionary [6]. The next study uses machine learning-based approach which uses morphological analyzer and probability of similarity and Hidden Markov Model (HMM) for candidates ranking [7]. The most recent one also uses machine learning-based approach which uses bigram/coocurrence/unigram model and HMM for candidates ranking [2]. These works gave a high accuracy. However, these works still require many manually defined rules. For other languages, there are spell checker studies that use deep learning-based approach [9] [10] [4]. However, no such work has been found yet for Indonesian language. We propose a Long Short-term Memory (LSTM) based encoder decoder model which examines word by word at character level that does not require any rules to correct spelling errors. We use words characters and context feature, in which the context features consist of adjacent words and Part of Speech (POS) tags. The adjacent words are represented by the word-embedding vector.