Developing Rule-based POS Tagger for Javanese: The First Stage

Totok Suhardijanto and Arawinda Dinakaramani

Abstract

This paper concerns our efforts to build a part of speech (POS) tagger for Javanese. This attemps is a part of our annotated modern Javanese corpus project. We divide our development of Javanese POS tagger into three different stages. The first stage is a study to analyze, design, and build Javanese POS tagset, disambiguation rules, closed-class taging dictionary, and multi-word expression (MWE) dictionary. The second stage is the development of morphological analyzer for Javanese, and the final stage is the creation of Javanese POS tagger.

In this paper, we only deal with the first stage because our attempt for developing Javanese POS tagger has just begun. This paper presents our research related to Javanese part of speech in our attempt to build a POS tagset and disambiguation rules for Javanese. For the time being, it is hard to find any research or textbooks which focuses on Javanese part of speech. The grammar book Tata Bahasa Jawa does also not provide us with a detail section which discusses a part of speech in Javanese. For this reason, we develop the Javanese tagset based on an existing Indonesian POS tagset. This paper also discusses our effort to build disambiguation rules for Javanese. The rules are expected to resolve the problem in disambiguate lexical items in Javanese.

Key words: POS tagging, Javanese, word disambiguation, computational linguistics, natural language processing