Semi-automatic Acquisition of Two-Level Morphological Rules for Iban Language

Suhaila Saee[1,2], Lay-Ki Soon[2], Tek Yong Lim[2],
Bali Ranaivo-Malançon[1], and Enya Kong Tang[3]

[1] Faculty of Computer Science and IT, Universiti Malaysia Sarawak,
Jalan Datuk Mohd Musa, 94300 Kota Samarahan, Sarawak, Malaysia
[2] Faculty of Computing and Informatics, Multimedia University,
Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia
[3] School of Computer Science and Information Technology,
Linton University College, Mantin, Negeri Sembilan, Malaysia

Abstract. We describe in this paper a semi-automatic acquisition of morphological rules for morphological analyser in the case of under-resourced language, which is Iban language. We modify ideas from previous automatic morphological rules acquisition approaches, where the input requirements has become constraints to develop the analyser for under-resourced language. This work introduces three main steps in acquiring the rules from the under-resourced language, which are morphological data acquisition, morphological information validation and morphological rules extraction. The experiment shows that this approach gives successful results with 0.76 of precision and 0.99 of recall. Our findings also suggest that the availability of linguistic references and the selection of assorted techniques for morphology analysis could lead to the design of the workflow. We believe this workflow will assist other researchers to build morphological analyser with the validated morphological rules for the under-resourced languages.

Keywords: Morphological rules, Rules extraction, Under-resourced language, Morphological analyser.

1 Introduction

Morphological analyser is a first processing task requires in Natural Language Processing (NLP). Morphological rules are crucial components in the analyser in order to analyse and generate the input word. The conventional method in acquiring the rules for morphological analyser was done using handcrafted, which has led to an ambiguity [1]. Therefore, the acquisition of the rules has received much attention from the researchers to automate the acquisition of the rules [2]. To automate the acquisition of the rules, there are two main components required as input: a) sufficient of linguistic references i.e. dictionary with stems and inflected words, the classification of words and affixes as well as a training data set and b) the selected techniques to acquire the rules that are depending

A. Gelbukh (Ed.): CICLing 2013, Part I, LNCS 7816, pp. 174–188, 2013.