Stemming Indonesian Words without a Dictionary

Vinsensius Berlian Vega[1], Stéphane Bressan[2]
[1]Genome Institute of Singapore
[2]National University of Singapore
vegav@gis.a-star.edu.sg, steph@nus.edu.sg

Abstract

The Indonesian Web is the set of digital documents in the Indonesian language available on the Internet and the World Wide Web and accessible from desktop computers, laptops, personal digital assistants and mobile phones. In order to provide the necessary seamless access to this wealth of information and to enable its development, the basic information retrieval tools need to be designed and deployed.

At the heart of a search engine retrieval mechanism for the English language is the stemmer. The stemmer fuses words that are morphological variants into the same root word. This obviously improves the retrieval performance in the case of inflectional variants. It is less clear in the case of derivational variants. There exists a number of stemming algorithms for English that do not require the word or its root to be checked against a dictionary. This is the case, for instance, for the Paice/Husk, Porter, Lovins, Dawson and Krovetz stemming algorithms.

The Indonesian language is morphologically rich. Some of the difficulties involved in the design of a practical stemmer for the Indonesian language are: the large number of affixes (prefixes, suffixes, circumfixes and even, but anecdotally, some infixes inherited from Javanese), the possible iterative use affixes, heavy usage of abbreviations and acronyms, loose standard for compound words, and rampant colloquial language, which include use of digits, use of obsolete and alternative spelling system, informal abbreviations, and emotive particles.

Our stemming algorithm does not rely on a dictionary unlike the algorithm by Nazief and Adriani, the Malay stemmer Ahmad, Yuso, and Sembok and the Malay morphological analyzer by Ranaivo-Malançon. We have developed several variants that accommodate or not colloquial language with emerging affixes (e.g. “n-”, “-in”) and that are more or less iterative. The algorithm is implemented as a Definite Clause Grammar in ECLiPSe Prolog. The algorithm looks up the available rules in order (e.g. “me-”, “di-“, “ber-”, “-el-”,”-em-”, ”-er-”, “-nya”, “-an”, “-i”), and greedily applies them. For example, “menghargai” is processed as “me”+”harga”+”i” and the root word “harga” is returned.

However, unlike French and Slovene, the morphology of the Indonesian language is more derivational than inflectional. We also present and discuss some results of experiments evaluating the impact of our stemming algorithm on the retrieval performance.