A Study on Indonesian Morphological Analysis

Hidayat Muizzul

School of Information Science,
Japan Advanced Institute of Science and Technology

May 15, 1997

Keywords: Indonesian, Morphological Analysis, Language Knowledge Acquisition.

Morphological analysis is one of the most important parts in natural language processing. It is a basic process that uses a dictionary to analyze a sentence and then provides lexical information about words in the sentence. Although there are a lot of researches on morphological analysis in many languages in the world, few researches are made on Indonesian morphology. Indonesian is a language which is used by nearly 200,000,000 peoples. It has several special characteristics that cannot be seen in other languages. While Indonesian looks like English in the way that it occupies a space between words in a sentence, many words may be devided into some small parts, including prefixes and suffixes. We call such words, compound words. On the other hand, we call a word that cannot be decomposed any more, basic words. A compound word is composed of some prefixes, a basic word and some suffixes. With a certain affix (prefix and suffix), the compound word will be a derivation word of the basic word but occupy some degree but not entirely different, such as an affix may change the function (category) of a basic word from noun to a compound word which is a verb or vice versa but retains meaning which is similar to the basic word. These phenomena are more complex than those occurred in English and Japanese because almost basic words in Indonesian can appended with any affix. In English, most of words that can be added with an affix are verb, adjective, adverb and noun while words in Indonesian can be attached with any prefix and suffix not only one time but several affixes can be attached at the same time. In Japanese, there is no space between words and the task in morphological analysis is to segment words from each other. Indonesian is different from Japanese is the way that a word can be attached with several affixes in the same time. There are not so many words that can be attached by affixes in Japanese. Moreover, one of the interesting phenomena in Indonesian is the duplication of a word. A new word can be formed from a repetition of a certain word.

Recently some Indonesian dictionaries have been built by research organizations. The representative dictionary is that provided by CICC. This dictionary includes nearly 31,000

words with syntactic, symantic tags and English translation. However, building a large-scale dictionary requires a lot of effort and labor and most hand-made dictionaries include only a limit set of words that are widely used in Indonesian.

In this paper, we propose a method to construct a wide-coverage dictionary from an existing dictionary with a limitted vocabulary (CICC Indonesian dictionary). Instead of recording all words (that is, all basic word and all compound words) in the dictionary, our method keeps only basic words and some rules and applies them to construct compound words from the basic words. To do this, a set of basic words and their compound words are extracted from the dictionary at the first place. The compound words are devided to their basic words, their affixes. In Indonesian, an affix is sometimes a derivative of its root form and maybe composed of more than two priemitive affix.

Applying some simple heuristics, primitive affixes and their root forms are acquired. As the result, the above-mentioned knowledge are later kept in basic word dictionary, prefix dictionnary, suffix dictionary and a database of rules for building compound words. The rules are in the form of connection matrix indicating that which prefix can be followed by which basic word, which basic word can be followed by which suffix, and so on. By this way, it is possible to reduce the size of dictionaries and also extend the existing dictionary to have more ability to analyse words which are unseen in the dictionary. Moreover, this thesis also provides the way to assign a category to the word in the morphological analysis. In principle, two any words in Indonesian tend to have a same category when they owe a same affix. This phenomenon was also reported by several Indonesian linguists. Due to this, we construct a rule table for this purpose and its accuracy is investigated by way of experiment.

Finally, we show the effectiveness of our system by some experiments. In the experiments, we use Indonesian corpus including around 1,000 sentences acquired from an Indonesian newspaper. One experiment is made to check how much our system can extend the coverage of a dictionary. Another experiment is done for checking the accuracy of categories assigned to words in a sentence.