IORA International Conference on Operation Research
Bogor, Indonesia, August 27, 2016

Affixed Word Classes Identification Through Stemming Process With Porter Stemmer Algorithm

Iyan Maulana[1*], Aries Maesya[2]

[1,2]Department of Computer Science, Universitas Pakuan, Indonesia

[*]corresponding author: iyandelon@yahoo.com, a.maesya@gmail.com

Abstract. In the application of information retrieval, stemming is an important process to accelerate indexing and query processes. Besides that, stemming process can also be used to determine the word class of basic or affixed words. This research is aimed at identifying the affixed words class before stemming and the basic word class after stemming. The stemming method applied is porter stemmer method with the steps of deleting the particles, deleting the pronouns, deleting prefix 1, deleting prefix 2, and the last is deleting suffix. The result shows that porter stemming still needs to be added by some other algorithms in order to improve the accuracy of stemming results in Bahasa Indonesia. However, the identification result of word classes both for basic and affixed words still shows 80% of accuracy and is still influenced by the stemming process conducted.

Keyword : information retrieval, stemming, word class, porter, stemmer

Introduction

Stemming process is a process of deleting affixes from a word so that the basic word is gained. [1] In the application is a process which is needed in accelerating indexing and query processes. It is because in information retrieval, affixes are part of information which are not meaningful, so it needs to be deleted. There are two ways of stemming that can be done using information retrieval application. The first way is using dictionaries by comparing and data mining from database. The second way is using stemmer algorithm which uses affix rules. [2]

In the similarity measurement of documents written in Bahasa Indonesia, the use of rules based algorithms showed high number of errors and it might influence the accuracy of the result. [3] However, the performance of rules based stemming is relatively stable with growing number of documents. It is because Bahasa Indonesia is a language having tendency of using affixes freely. At least there are 35 official affixes mentioned in *Kamus Besar Bahasa Indonesia* (Big Dictionary of Bahasa Indonesia). The affixes include prefix, suffix, and infix. [4]

To enhance the quality of temu kembali informasi, stemming process is not only used to gain the basic word but also to determine the word class of an affixed word. It is because each word class has its own meaning. A basic word followed by an affix will build a different meaning from its basic word. Through word class identification, it is expected that it can help people choose the right words.

Stemming Technique

Stemming Technique in Bahasa Indonesia includes various methods. The first method is stemming with affix cutting reference table. The stemming process of a term with this method is conducted by cutting the affixes of the term according to the table. The second method is the development of the first one. The second method uses not only affix cutting reference table but also a dictionary containing basic words. The dictionary is used as the reference when affix cutting has been done. The result of stemming will be correct when the basic word exists in the dictionary, but if it is not, then the input word will be considered as the basic word. [5]

An example of determining word class and the morphology of an affixed word which is "memakan" as an input is shown as follows:

1. Input the word "memakan"
2. The affix determining:
   a. Particles, such as *lah, kah,* and *pun*,
   b. Pronouns, such as *ku, mu,* and *nya*,
   c. Prefixes 1, such as *peng, meng, peny, meny, pen, men, pem, mem, pe, me, di, ter, and ke*,
   d. Prefixes 2, such as *ber, bel, be, per, pel, pe, and se*,
   e. Suffixes, such as *kan, an,*and *i*
3. Do the stemming process,
   a. Check the particles. If there is, then delete the particle of the word and save the particle in the new variable, so that the word "memakan" becomes "memakan",
   b. Check the pronouns. If there is, delete the pronouns of the word, and save it in the new variable, so that the word "memakan" becomes "memakan",
   c. Check prefix number 1. If there is, then delete the prefix number 1 of the word and save it in the new variable, so that the word "memakan" becomes "makan" and prefix number 1 is "me",
   d. Check prefix number 2. If there is, then delete the prefix of the word and save it in the new variable, so that the word "memakan" becomes "makan",
   e. Check the suffix. If there is, then delete the suffix of the word and save it in the new variable, so that the word "memakan" becomes "makan",
4. The basic word is "makan", check the database. If there is, then take the field of basic word of "makan", and the class is verb and an affix.
5. Declare it to the data variable in number 4. A= "makan", b = verb, prefix 1 = "me", ,
6. If prefix 1 = "me" is followed by a = "makan", the affixed word class is verb,
7. Declare the affixed word class = verb into new variable,
1. If the class = verb and b = verb, then the category of morphology is verba deverba.

Research Steps

Generally, this system has two research steps. The first system will process affixed word input to be basic word. This process is called as stemming and the method used is porter stemmer. The second step is determining the class of the affixed word and determining the morphological category from the input affixed word.

The first picture shows the data flow from the method of porter stemmer. The affixed word which has been input will be processed into the basic word. The basic word will be completed with its word class. The result is important since it will determine the next process which is the determining of the affixed word class and the determining of the class morphological category.

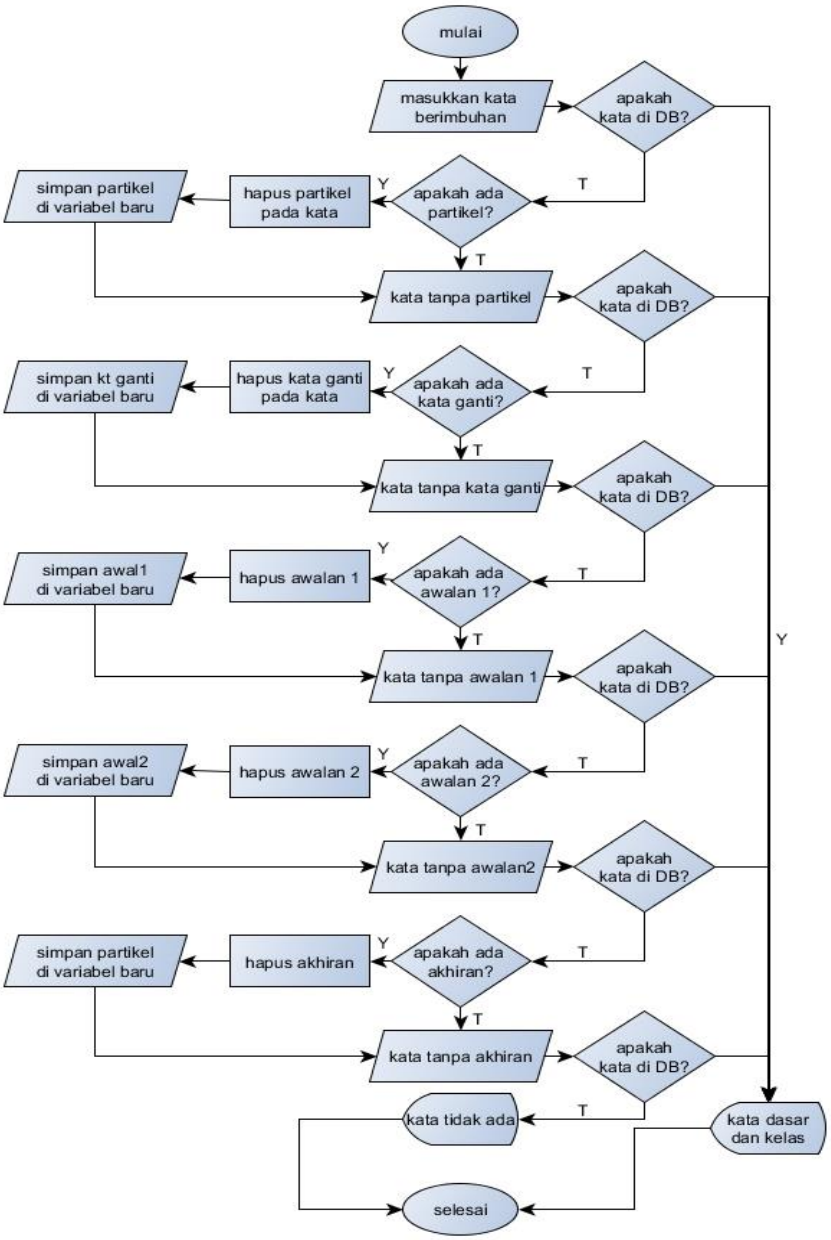

Figure 1. ***Flowchart Porter Stemmer***

In the next step, the word class and the morphological category will be determined like it is shown by figure 2

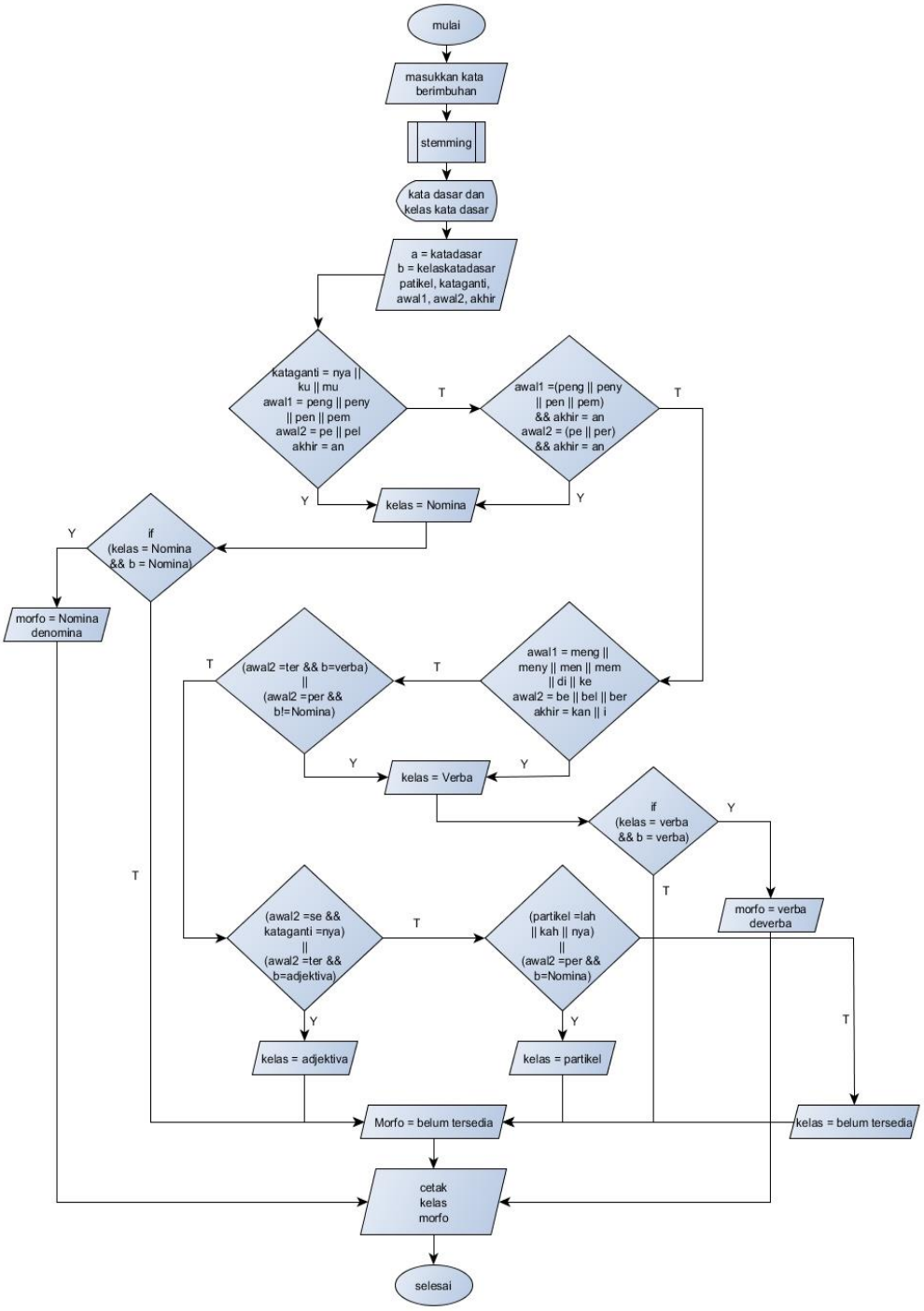

Figure 2. ***Flowchart Determination of Class Affixation And Categories Morphology***

Research Result

The identification of affixed word class is started by stemming process to delete its affixes. In this research, porter stemmer method is used with the steps of deleting the particles, deleting the pronouns, deleting prefix 1, deleting prefix 2, and finally deleting the suffix.

The result of stemming with porter stemmer algorithms at the beginning of the research shows low level of accuracy because stemming process error still occurred and it influenced the identification result of the affixed and the basic word classes. As an example, the word "memasak" which the stemming result was supposed to be "masak" yet it appeared "pasak". Thus in this research, some new algorithms are added so that the result is "masak". The following is the comparison between the previous research result and the new one:

Table 1. The Comparison of Stemming Porter Method

Hasil Sebelumnya		**Hasil Penelitian terbaru**	
Inputan 1	Memasak	Inputan 1	memasak
Hapus Partikel	Memasak	Hapus Partikel	memasak
Hapus Kata Ganti	Memasak	Hapus Kata Ganti	memasak
Hapus Awalan 1	Pasak	Hapus Awalan 1	masak
Hapus Awalan 2	Pasak	Hapus Awalan 2	masak
Hapus Akhiran	Pasak	Hapus Akhiran	masak
Kata Dasar	Pasak	Kata Dasar	masak
Inputan 2	memandang	Inputan 2	memandang
Hapus Partikel	memandang	Hapus Partikel	memandang
Hapus Kata Ganti	memandang	Hapus Kata Ganti	memandang
Hapus Awalan 1	pandang	Hapus Awalan 1	Pandang
Hapus Awalan 2	pandang	Hapus Awalan 2	Pandang
Hapus Akhiran	pandang	Hapus Akhiran	Pandang
Kata Dasar	pandang	Kata Dasar	pandang

The identification process of word class is done to affixed word class which is before stemming and the basic word class after stemming. The identification of word class used is dictionary method which compares and takes the data available at the database. The identification result of basic word class and affixed word class is shown by Picture 3 with the average of accuracy of 80%.

Inputan	imbuhan
kelas kata	kata imbuhan merupakan kelas kata nomina
Kategori	kata imbuhan merupakan Kategori Morfologi Nomina Denomina
Kata Dasar	imbuh adalah kelas Nomina

Figure 3. The Result of Basic and Affixed Word Classes