Rule-based Part of Speech Tagger for Indonesian Language

K K Purnamasari1* and I S Suwardi2

1Faculty of Engineering and Computer Science, Universitas Komputer Indonesia,
Jalan Dipatiukur No. 112-116, Bandung, Indonesia
2 School of Electrical Engineering and Informatics, Institut Teknologi Bandung,
Jalan Ganesha 10, Bandung, Indonesia

*ken.kinanti@email.unikom.ac.id

Abstract. Lexical processing, such as detection of root words (Stemming) and type of words (Part of Speech tagging) is the important determinant for language computing systems that requires the detection of sentence structure or pattern. In Indonesian language, a problem that was encountered in lexical processing is lack of annotated corpus in Indonesian. Thus, POS tagger built in this research did not use annotated corpus, but utilize KBBI (Indonesian large dictionary) and some morphological rules that affect changes in word form (morphology). The method used in this study begin with change input text into tokens and do stemming. Each token is checked whether it is available in the baseword dictionary or not. If the token is not found in the baseword dictionary, it will go through stemming and affix detection. By doing this step, it can produce output list of POS tag for basewords and their affixes. By collate the output based on the rules of grammar, we can determine the type of affixed-word. Testing is done by comparing the detection results and the type of wordlist available in KBBI, for every input token. Accuracy score is obtained by calculate number of true results divided by total number of tokens being examined. Based on result of test performed, the achievement of accuracy is quite good (average rate of 87.4% for 4 parts of PAN Localization corpus in Indonesian). False results were caused by some mistake tags in existing KBBI and presence of ambiguous word (word with more than one POS tag). So, improvement will be possible by using more complete Indonesian dictionary and adding word-sense disambiguation.

1. Introduction
Proper preprocessing tools in many Natural Language Processing (NLP) studies are very important to give better accuracy. Lexical preprocessing stage, such as detection of root words (Stemming) and the detection of type of words (POS tagging) have great impact for language computing systems that require the determination of sentence structure. In Indonesian language, research on Stemming and POS tagging is still being conducted, either by using statistical methods or certain rules. Some problems were encountered for such processing is the lack of corpus in Indonesian and the incompleteness of available rules. Research on stemming, first published by Julie Beth Lovins in 1968 [1]. Followed by Martin Porter, twelve years later in a journal [2]. Currently, there are several methods of cutting in stemming algorithms for Indonesian language, which are Nazief and Andriani [3], Arifin and Setiono [4], Vega [5] and Tala [6]. Result from an experiment comparing the algorithms in Table 1, show that Nazief-Andriani algorithm give highest accuracy [7]. Thus, this study will use Nazief-Andriani algorithm to do stemming with some modifications tailored to handle the specific cases.
Meanwhile, construction of POS Tag, has previously performed probabilistic [8][9], rule-based [10], and the incorporation of both [11]. Wicaksono algorithm use Hidden Markov Model (HMM) with 35 POS tags. This research resulted in an accuracy approaching 99.4% without Out of Vocabulary (OOV) handling. While Fam Rashel implement Conditional Random Field (CRF) and Maximum Entropy (ME) with 37 and 25 POS tags. In both probabilistic solutions, accuracy level will continue to change depend on quality of selected corpus. The limited number of adequate Indonesian annotated-corpus make it impossible to achieve significant improve in probabilistic way (Table 1).

Table 1. Comparation of Stemming Algorithm.

| Algorithm                         | Advantage                     | Lack                          |
|-----------------------------------|-------------------------------|-------------------------------|
| Nazief-Andriani                   | Good accuracy                 | Large memory                  |
| (with dictionary and recoding)    |                               |                               |
| Arifin-Setiono                    | Overstemming could be         | Prefix and Suffix             |
| (with dictionary)                 | handled by affix combination  |                               |
| Vega Bressan                      | No need dictionary            | Low accuracy                  |
| (without dictionary)              |                               |                               |
| Ahmad, Yusoff, Sembok            | Overstemming could be         | Accuracy depends              |
| (with dictionary and recoding)    | handled                       | on rule’s order               |
|                                   |                               |                               |

So, this study purpose to built tagger pos use another path by utilizing the concept of using Indonesian dictionary with some rules created based on Indonesian grammar. There will be a slightly different approach in the implementation of stemming process, because the affixes which are usually discarded will be used as the main sources in POS tag detection. The attention for changes in word form (morphology) is expected to improve the accuracy obtained.

2. Methods
Previous studies about POS Tagging utilized statistical methods such as Hidden Markov Model (HMM), Conditional Random Field (CRF), and rule-based methods such as Maximum Entropy (ME). These methods require an annotated corpus as its main input, both as the model (in statistics) as well as the rule (the rule-based). This causes the accuracy of output produced depends on the quality of the corpus used. Another obstacle is incompleteness and the lack of resources quality in Indonesian language corpus. In this study, annotated corpus only be used in the scope of testing. Rules are made in accordance with morphological pattern that can be known from the words in the Indonesian language in general. The detection of the type of words in this study conducted by utilizing particle detection in the process of stemming. The entire process built is divided into three main parts : Tokenization, Affix Detection and POS Tagger (Figure 1).

Figure 1. POS Tagging Stages.

2.1. Tokenization
A sentence in a text consists of a series of words that begin with a capital letter and ends with a period, question mark, or exclamation [12]. Based on this condition, the text can be processed into a collection of tokens. In this work, tokens can be words, symbols, or numbers.

2.2. Affix Detection
Affix detection is done by applied roots of words detection (stemming). This work utilized Nazief and Andriani stemming algorithm approach, because it has the highest accuracy compared with three other
methods [4][5][13]. Some modifications are taken to make it suitable for POS tag detection. Using the algorithm, the system can produce root word of each token detected. By separating the root word, it can show the existence of affixes (prefix, suffix, insertion, or melt particle) from each token.

2.3. Rule-based POS Tagging

POS tag classification in this study utilized the detection result of root words and affixes in the previous stage. First, we need to get the POS tag of each root word (from root words dictionary). Then, employ some morphological rules built upon the presence of affixes in each token. These will give the suitable POS tag for each token as the result.

3. Results and Discussion

POS tagging implementation, conducted by utilizing morphological rules for affixed words. Particle detection process begins with the tokenization which process the affixed words in a text. Each of them is broken down based on their root word and affixes. This system is divided into 4 main functions: Text Input, Tokenization, Parse Dictionary & Check Affix, and Testing. Text Input, retrieve data by doing a search on the local and online documents (by entering the desired URL). Tokenization, separate the text into tokens. Parse Dictionary & Check Affix, check out the tokens that have been generated by the previous function. If a token that is checked in the dictionary, it will produce the appropriate tag type of word in the dictionary. If not, it will be examined the rules change kind words from the base word to affixed words. Testing, compare the results of the detection was done with words that have been defined on KBBI (Indonesian Dictionary). All three are implemented in these execution stages: Text Input, Tokenization, and Classification (Dictionary parsing, Affix detection, and Testing). Based on experiments that have been carried out, the accuracy percentage will be calculated by the following formula. The results can be seen in Table 2.

| Input text (.txt)               | Total (words) | Accuracy (%) |
|---------------------------------|---------------|--------------|
| PANL_bppt_economy_01_1017       | 1017          | 90           |
| PANL_bppt_economy_02_4287       | 4287          | 91           |
| PANL_bppt_economy_03_1615       | 1615          | 88           |
| PANL_bppt_economy_04_1894       | 1894          | 87           |
| PANL_bppt_economy_05_1235       | 1235          | 86           |
| PANL_bppt_international_01_1201 | 1201          | 86           |
| PANL_bppt_international_02_2025 | 2025          | 86           |
| PANL_bppt_international_03_2124 | 2124          | 87           |
| PANL_bppt_international_04_1485 | 1485          | 85           |
| PANL_bppt_international_05_4338 | 4348          | 88           |
| PANL_bppt_science_01_1101       | 1101          | 85           |
| PANL_bppt_science_02_2012       | 2012          | 87           |
| PANL_bppt_science_03_4275       | 4275          | 87           |
| PANL_bppt_science_04_1369       | 1369          | 88           |
| PANL_bppt_science_05_1158       | 1158          | 87           |
| PANL_bppt_sport_01_1325         | 1325          | 88           |
| PANL_bppt_sport_02_1816         | 1816          | 88           |
| PANL_bppt_sport_03_1968         | 1968          | 88           |
| PANL_bppt_sport_04_4506         | 4506          | 88           |
| PANL_bppt_sport_05_1235         | 1235          | 88           |

Average                                           87.4
4. Conclusions
Although it produces fairly good accuracy (87.4%), POS tag detection result still has not been able to deal with ambiguous words (word with more than one POS tag). Thus, a development of this research can be improved by adding word-sense disambiguation and examining morphology pattern of slang words, so it can be used to recognize POS tag from the higher diversity corpus. It can also add more specific terms to handle specific scope of corpus (for example: biological terms, politics, engineering, and others).

References
[1] Lovins J B 1968 Development of a stemming algorithm Mechanical Translation and Computational Linguistics 11 pp.22–31
[2] Porter M F 1980 An algorithm for suffix stripping Cambridge
[3] Adriani M, Asian J, Nazief B, Tahaghoghi S M M and Williams H E 2007 Stemming Indonesian : A confix-stripping Approach ACM Transactions on Asian Language Information Processing (TALIP) 6
[4] Ariffin A Z and Novan A N 2002 Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algoritma Single Pass Clustering Proceeding Seminar on Intelligent Technology and its Applications (SITIA)
[5] Vega V B 2001 Information Retrieval for the Indonesian Language Master Thesis
[6] Tala F Z 2003 A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia Master Thesis
[7] Asian J, Williams H E and Tahaghoghi S M M 2005 Stemming Indonesian Conferences in Research and Practice in Information Technology Series 38 pp.307–14
[8] Pisceldo F, Adriani M and Manurung R 2009 Probabilistic Part of Speech Tagging for Bahasa Indonesia Proceedings of 3rd International MALINDO (Malay-Indonesian Language) Workshop
[9] Wicaksono A F and Purwarianti A 2010 HMM Based Part-of-Speech Tagger for Bahasa Indonesia Proceedesings of 4th International MALINDO (Malay-Indonesian Language) Workshop
[10] Rashel F, Luthfi A, Dinakaramani A and Manurung R 2014 Building an Indonesian Rule-based part-of-speech tagger Proceedings of the International Conference on Asian Language Processing
[11] Widhiyanti K and Agus H 2012 POS Tagging Bahasa Indonesia Dengan HMM dan Rule Based Informatika 8 pp.132-156
[12] Sakri A 1994 Ejaan Bahasa Indonesia (Bandung: Institut Teknologi Bandung)
[13] Fathony A, Yusoff M and Sembok T M T 1996 Experiments with a Stemming Algorithm for Malay Words Journal of American Society for Information Science 47 pp. 56-72