DEVELOPMENT OF DETECTION INDONESIAN PARTS-OF-SPEECH DICTIONARY AS NLP RESOURCE

Arief Adiguna Putra[1], Ken Kinanti Purnamasari[2]

[1,2]Informatics Engineering – Indonesian Computer University
Dipatiukur Street No 114-116 Bandung 40132
E-mail : ariefadiguna@gmail.com[1], guinev_gs@yahoo.co.nz[2]

ABSTRACT

A dictionary is a book containing the words along with a description of its meaning, its use or its translation. Along with the development of the times, any dictionary can be found in digital form that can be accessed online. In its development type of word dictionary has been constructed with data from Indonesian Dictionary (KBBI). In the Indonesian Dictionary (KBBI) has a number of words as 65.857 entries consisting of the principal amount of entry (the root word) as many as 41.250 and derivative as 24.607 words as many entries. Development of Indonesian part-of-speech dictionary as NLP resource has got the result as much as 38,870 entry [1]***. Problems in previous research that affects the amount of the entry is detected is input data conversion techniques are used and the steps retrieval of words and kind words are still less effective.***

From the results of the analysis that has been done, then consists of the process of smoothing the data input which aims to spruce up the input data and the process of detection of words and type of words which aim to detect words and kind words in accordance with the format set composed of words and type of words.

The results of this research have been getting words 43,060 as much with the percentage of 65.38% of the total number of root words and its derivatives as much as 65,857 words. Based on the analysis, the increase in detection result occurs because of the optimization of the technique per alphabetical input data conversion and settlement of problems at the stage of making root words and derivatives.

Keywords: Dictionary of type words, online dictionary, word-class classification, data sources NLP, WordNet.

1. INTRODUCTION

Type of word is a word class in language units by category forms the function and meaning in the makeup of the grammatical system. To construct a sentence that is good and right with the sentence on the patterns of raw, language users must know the type and function of speech. Dictionary types of words has been implemented on several studies in the field of NLP (Natural Language Processing) as the classification of types of words in the POS Tag, Stemming, Development WordNet and others.

Along with the times, the dictionary can be found in digital form that can be accessed online. In dictionary online is not accompanied by resources, in other words, users simply wear and does not have access rights to change or add. In the Indonesian Dictionary (KBBI) has a number of words as 65.857 entries consisting of the principal amount of entry (the root word) as many as 41.250 and derivative as 24.607 words as many entries.

In development of Indonesian part-of-speech dictionary as NLP resource has got the result as much as 38,870 entry [1]. Problems in previous research that affects the amount of the entry is detected is input data conversion techniques are used and the steps retrieval of words and kind words are still less effective.

So in this research will be the analysis of the sources of input data and at each step by step in the detection process words and types of words, with it will be known deficiencies that affect the number of entries gained. So that the optimization process will be conducted in detecting words and types of words that are not detected and will modify input data conversion technique that was originally located in one file is broken down into alphabetical.

2. RESEARCH CONTENT

Formulation of the problem according to the above background is not detected all words (entries) in the Big Indonesian Dictionary (KBBI) in previous research.

2.2 Dictionary

Dictionary in Arabic, namely Qamus (plural qawamus). Arabic absorb the word dictionary of words in ancient Greece, Okeanos which means sea. Classification is based on the language dictionaries that can be written in one or more than one language is composed of several types of eka language dictionaries, bilingual dictionaries and a multilingual dictionary. Development resource dictionary to be built type of words included in the dictionary eka language, in which words are processed using only one language. The word is a collection of some of the letters that have specific meanings. In a word is an element of spoken or written language that embodies a feeling and thoughts that can be used in the language. [2] Type of word is a word class in language units by category forms the function and meaning in the makeup of the grammatical system [**2**]. The classes of said resource development type of words used as a reference is based Grammar Indonesian with a commonality in the classification of speech as in the Indonesian Dictionary who became a source of input data.

2.3 Analysis Of The Problem

Problems are taken based on the background discussed above is the detection of the type of words are studied in development of Indonesian part-of-speech dictionary as NLP resource has got the result as much as 38,870 entry with the total number of entries in the Indonesian Dictionary (KBBI) which has a number of root words and derivatives many as 65 857 entries consisting of the principal amount of entry (the root word) as many as 41 250 and 24 607 derivatives words as many entries.

In previous studies the data input of Indonesian Dictionary (KBBI) are in a plain text file format (.txt) and the main process is the collection phase word and a kind word and a conversion output format. Stages making process words and kind words consist of ten stages, which consists of the removal process diagan digit lines that are less than two words, making words and types of words, check words in line, and sequencing. Based on the analysis in each process, the cause of many entry (words) is missing occurs in the input data and the retrieval process words and type of words. [**1**] At the stage of making words and kinds of words in previous studies, no categorization of stages of pre-processing and primary processes. Here are the stages of making words and kinds of words in previous research.

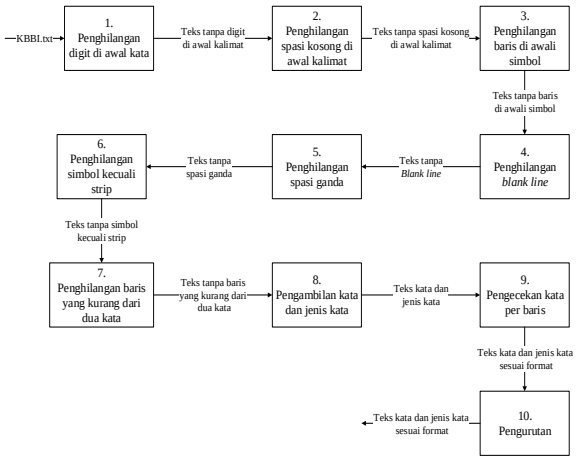


Figure 1. Stages Taking Words Of Previous Research

From the stage of taking words and type of words following entries have been able to detect as many as 38 870 entries with the percentage as much as 59.01%. The results of the comparison stages of decision-word can be seen in the following table.

Table 1. Comparison Of Process

No	Stage Taking Words On Previous Stage	Process Result
1	Elimination of Digits in the Begin	-
2	Elimination of Empty Space in Early Sentence	-
3	The row deletion begins Symbol	-
4	Delete Blank Line	-
5	Delete Double Space	-
6	Delete Symbol Unless Strip	-
7	Delete Missing from the line 2 Words	-
8	Taking Words and Type Of Words	36.988 Line/Words
9	Checking Words Per Line	38.870 Line/Words
10	Sorting	38.870 Line/Words
Detections Result		38.870 Line/Words
Persentase		59,01%

2.4 Analysis Solutions

Analysis of the solution is a stage where the process is to identify any problems and the needs required to make the experiment by other means. The solution provided in this study were influential in increasing the number of entry is the data conversion process will be three (3) phases of the original .pdf format will change data format to .doc and last change to .txt. The conversion process will be done alfabetically or one by one from the alphabet A to Z. Then there will be optimized in step retrieval word and a kind word so the problem - the problem can be solved.

In the analysis of solutions / systems to be run have clean up phase detection input data and root words and derivatives. Explanations of each stage can be seen in the picture below.

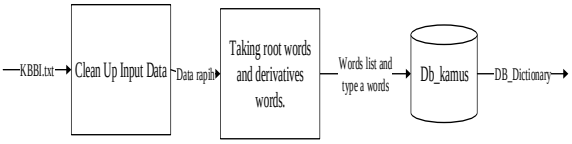


Figure 2. System Architecture Overview

On Architecture Overview The system will be explained as follows.

1. In the initial phase the user enters a file which is an Indonesian dictionary format plain text (.txt) into the system. 2. At this stage of smoothing the input data deletion process will be conducted a number of meanings, equating the separation distance between letters, deleting rows of the compound, the elimination of punctuation symbols except hyphens, deletion of rows is less than 2 (two) words.
2. In the detection stage base word and its derivatives will be done making process words and types of words, the separation line of each entry, duplicate checking, sorting.
3. The results of the phase detection basic words and kinds of words in the form of a list of words and kinds of words with the data format plain text (.txt).
4. Next, list the word format and types of words that are the result of taking words and types of words are stored into the database. In this case the database used is MySQL.

2.5 Analysis Of Data Input

Analysis of the data input to the detection system is kind word dictionary that describes the process of data input of text file data is received and will be processed by the system.

2.6 Conversion Process Data Input

The process of initial data format conversion .pdf format will be done in two phases, namely:

The process of changing the initial data sources within a single file into a alphabetically in .doc format and to format conversion process is done manually using Nitro Pro 10 software.

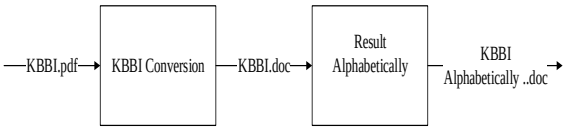


Figure 3. Process Conversion Data To .doc

Here is an example before the conversion process that can be seen in the following figure.

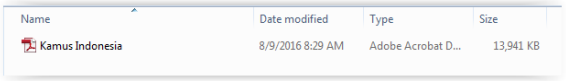


Figure 4. Data KBBI Before Conversion

Here is an example after the conversion process is done manually using Nitro Pro 10 software, such as file conversion process results KBBI alfabetically of the alphabet A-Z with .doc format that can be seen in the following figure.

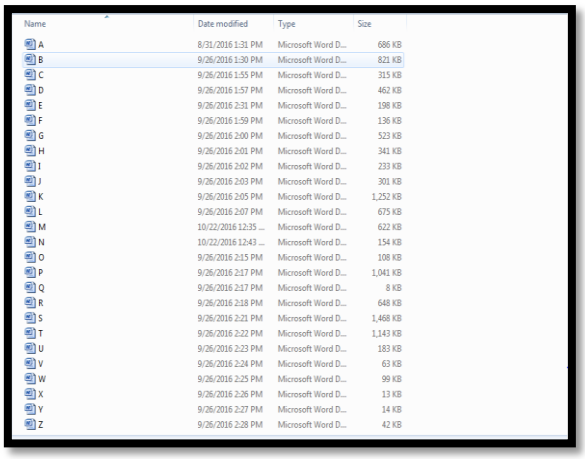


Figure **5.** Data KBBI Conversion To .doc

1. The results of the data dictionary alfabetically with .doc format it will be converted to plain text (.txt) and for format conversion process is done manually with Nitro Pro 10 software.

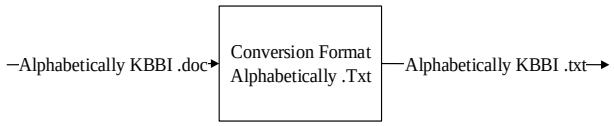


Figure 6. Comversion DataTo *.txt*

Here is an example before the conversion process that can be seen in the following figure.

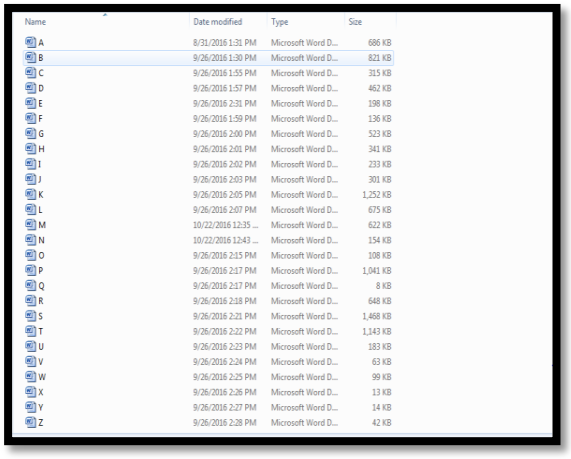


Figure 7. Data KBBI Before Conversion To .txt

Here is an example after the conversion process for KBBI alfabetically with plain text format (.txt) which can be seen in the following figure.

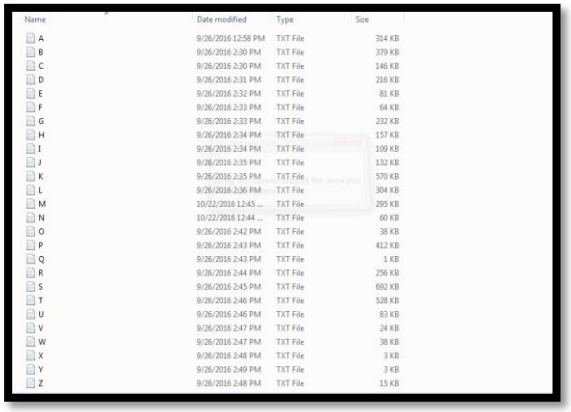


Figure 8. Data KBBI After Convesion To .txt

2.7 Data Input

Data input in the form of a kind word detection dictionary word text data and kind words alfabetically obtained from Indonesian dictionary with plain text format (.txt). Examples of input data to be processed can be seen in the following table.

Table 2. Data Input

1A, a n huruf pertama abjad menghablur v menjadi hablur menghablurkan v menjadikan hablur habuk n benda yg halus spt tepung debu a n are Indonesia 2A n Ampere; lambang satuan ukuran arus listrik a n are
3a n are; nama satuan ukuran 1adang v, mengadang v 1 menghalangi bisa luas (= 100 m2) a n are ab ark n tabung atau kotak candu terbuat dr tanah aba n ayah; bapak; (kadang-kadang juga berarti) kakek abadi a kekal selama-lamanya; tetap ada sepanjang masa; tidak berkesudahan: ha- nya Tuhan saja yg --; semua orang men- dambakan kedamaian yg --; aba-aba n kata-kata perintah atau ko- mando, spt dl baris-berbaris, senam, atau menyanyi bersama, msl siap!, kiri! kanan!, satu! dua!, maju jalan! abad n 1 masa seratus tahun: umurnya su- dah setengah --; 2 jangka waktu yg la- manya seratus tahun: -- ke-20 dimulai dr tahun 1901 sampai tahun 2000; 3 zaman (yg lamanya tidak tentu); 4 kl masa yg kekal, tidak berkesudahan; -- kekal selama-lamanya; -- keemasan masa kegemilangan dan kejayaan yg di- alami suatu bangsa atau negara dl se- jarahnya; -- pertengahan zaman dl se- jarah Eropa antara zaman purbakala dan zaman baru (tahun 476--1492); berabad-abad v beberapa abad lamanya; beratus-ratus tahun: ~ lamanya bangsa kita dijajah; seabad n satu abad; seratus tahun 2abuk n rambut: -- kaki; -- nya kusut masai semua 3abuk v, mengabuk mengambil atau meng- akui barang orang lain sbg miliknya abuk-abuk n nama kue yg dibuat dr tepung aren; kue abuk abulhayat n bapak kehidupan; hujan aba n ayah; bapak; (kadang-kadang juga berarti) kakek abun-abun ark n angan-angan;

2.8 Analysis Process

The analysis process consists of smoothing the process of data input and retrieval process words and kind words. Stage process of smoothing the input data consisting of 6 (six) stage of the process of smoothing the input data consists of the elimination of a number of meanings, equating the separation distance between letters, deleting rows of the compound, the removal of empty lines, the

elimination of punctuation except the hyphen, and deleting rows is less than 2 (two) words.

In the process of making words and kinds of words made up of 4 (four) process which is taking words and types of words, the separation line of each entry, checking duplication, and sequencing. Data input is the data sourced from Indonesian dictionary has been done so that the format conversion process to produce data dictionaries alfabetically with plain text format (.txt). The stage can be seen in the following figure.

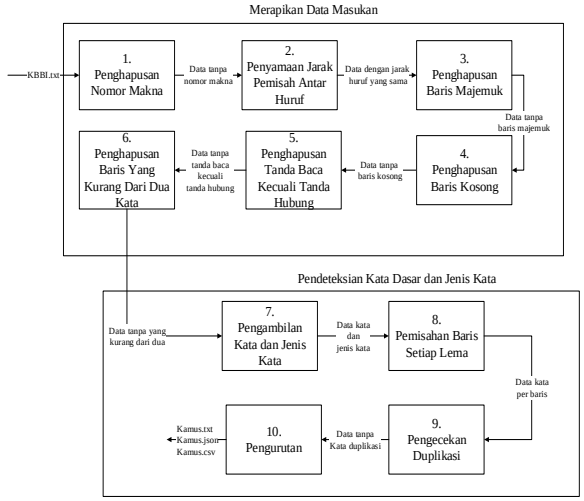


Figure. Process Scheme

2.8.1 Detection Process Basic Words

In making stage and kind words said aims to detect words and kinds of words in accordance with the format. Stages of making words and kind words do to get a sentence structure that consists of words and kinds of words (example: could n). Data inputs are processed separately alfabetically of the alphabet A to Z is derived from Indonesian dictionary Plain Text format (.txt). Here is a step-making process words and kind words.

1. Taking Basics Words

This stage is the retrieval word by word specified structure. The purpose of this phase is to take the word and types according to a predetermined format, starting with finding words and kinds of words per line, for example: could n. At the stage of making words and kind words earlier research, the process of making words and kind words could just take a word at the beginning of the line and can not take a word that is not followed by a kind word. Therefore, in this study will be optimized so that the problem - the problem can be solved. So that the results will improve its detection. For more details depicted in the block diagram below.

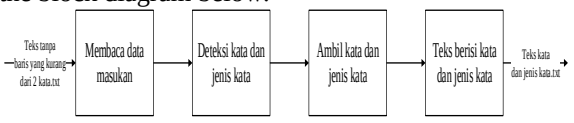


Figure 16. Block Diagram Intake Basics Words

The next process is taking words and kind words. The first stage begins with reading the data from previous stages. After that the replacement label words with spaces for the label word is not used in this detection. Label the word in question is the ark, cak, ho, cash, and kl. Here is an example of the replacement of the word label with spaces that can be seen in the following figure.

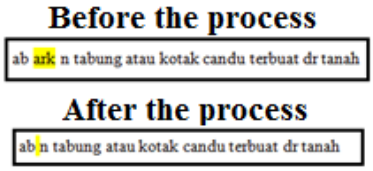


Figure **17.** Replacement Word Label

The next process is the search word and a kind word in each line, the search process based on predetermined rules that search within each row containing the letters az with at least one letter followed by a space and followed by the symbol of the kind of words that have been determined to be n, v, a, adv, num, p, or pron then if there was corresponding with these rules shall be regarded words and kind words. The process will be carried out repeatedly - again and will stop after not found words that are not followed by the kind of words on each line. Search results then saved into a .txt file that contains the data dictionary. As an example in the table below.

Table 3. Taking Basics Words

Data Sebelum Proses
A a n huruf pertama abjad
a n are
A n Ampere lambang satuan ukuran arus listrik
a n are
a n are nama satuan ukuran
luas 100 m2
a n are
ab ark n tabung atau kotak candu terbuat dr tanah
aba n ayah bapak kadang-kadang juga berarti kakek
aba-aba n kata-kata perintah atau ko-
mando spt dl baris-berbaris senam atau menyanyi bersama msl siap kiri kanan satu dua maju jalan
abad n 1 masa seratus tahun umurnya su- dah setengah -- 2 jangka waktu yg la- manya seratus tahun -- ke-20 dimulai dr tahun 1901 sampai tahun 2000 3 zaman yg lamanya tidak tentu 4 kl masa yg kekal tidak berkesudahan
berabad-abad v beberapa abad lamanya beratus-ratus tahun lamanya bangsa kita dijajah
seabad n satu abad seratus tahun
abadi a kekal selama-lamanya tetap ada sepanjang masa tidak berkesudahan ha- nya Tuhan saja yg -- semua orang men-
dambakan kedamaian yg --
adang v mengadang v 1 menghalangi

abuk n rambut -- kaki -- nya kusut masai abuk v mengabuk mengambil atau meng- akui barang orang lain sbg miliknya abuk-abuk n nama kue yg dibuat dr tepung aren kue abuk menghablur v menjadi hablur menghablurkan v menjadikan hablur habuk n benda yg halus spt tepung debu abulhayat n bapak kehidupan hujan aba n ayah bapak kadang-kadang juga berarti kakek abun-abun ark n angan-angan
Data Sesudah Proses
A a n a n A n a n a n a n ab n aba n aba-aba n abad n berabad-abad v seabad n abadi a adang v mengadang v abuk n abuk v mengabuk v abuk-abuk n menghablur v menghablurkan v habuk n abulhayat n aba n abun-abun n

2.8.2 Output Format Conversion

The results of some of the words making process and kind words with plain text format (.txt) file will be saved into a database that generates .sql format, then the next will be converted into a number of formats such as; .txt, .json, and .csv.

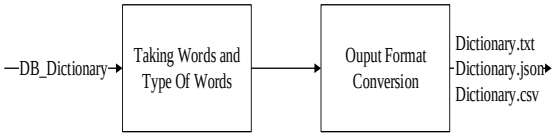


Figure 18. Block Diagram Output Format Conversion

2.9 System Test

System testing phase aims to find errors - errors or flaws - flaws in the system tested. The test aims to determine whether the program is a dictionary word and kind words made in accordance with the purpose of research.

2.9.1 Detection Test Result

In testing the accuracy will be calculated from the number of entries accuracy value that is obtained when the process of taking words and kind words. Number of entries gained as much as 43 060 entries on the total number of entries in the Big Indonesian Dictionary (KBBI) as many as 90 049 entries with the number of entries the principal (the root word) as many as 41 250 entries, the number of entry berimbuhan (descendent) of 24 607, and combinations of words (plural) of 23 536 entries. Then the results of the calculation as follows.

The total amount of principal and berimbuhan entries: 65 857

Total entries are available: 43 060

Detection Testing Results: = (Total Lema The Acquired / Total Lema KBBI) * 100% = (43,060 / 65,857) * 100% = 65.38%

So in this study resulted in the value of accuracy of 65.38%, with a yield peningkata many as 4190 entries (words) with a percentage of 6.36%.

Here are the results (outputs) of each phase detection process words and kind words.

Table 10. Table Test Result

Process Stage On Research	Process Result
Delete Number Meaning	-
Match Distance Separation Between Letters	-
Row Deletion Compound	-
Detele Blank Line	-
Except Deletion Punctuation Hyphen	-
The Row Deletion Less Than 2 Words	-
Taking Words and Type Of Worfds	40.878 Line/Words
Separation Line Every Words	46.184 Line/Words
Checks Duplication	45.316 Line/Words
Sorting	45.316 Line/Words
Detections Results	43.060 Line/Words
Persentase	65,38%

2.9.2 Analysis Of Results Testing

The following is an analysis of the results of detection test results that have been done. Based on the calculation basis of the word and its derivatives are done manually in the input data sourced from the alphabet A in Indonesian Dictionary (KBBI). Then obtained a number of basic words as much as 1774 entries and its derivatives as many as 853 entries. So the total entries gained as much as 2627 entries.

The comparison between the calculations manually at KBBI with the results of the detection process on the program can be seen in the following table.

Table 11. Comparison Results Detections

Manual Calculations	Detecions In Program
2627	2544

Based on the results of the comparison then there are 83 entries that can not be detected. From the results of the detection process, the compound / combination of words can not be detected as shown in the following figure.

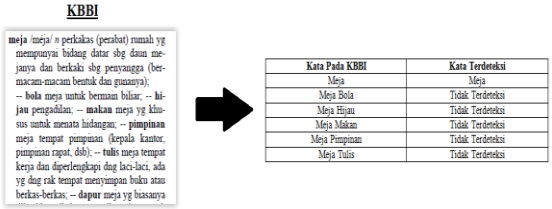


Figure 19. Analysis Of Results Testing

3. CONCLUSION

This section will explain the conclusions and recommendations of the research conducted.

3.2 Conclusions

In this case the development of detection dictionary to get the kind of words have as many entries as many as 43 060 entries with the percentage of 65.38% of the total number of basic words and derivatives as many as 65 857 entries. Based on the analysis, the increase in detection result occurs because of the optimization of the technique per alphabetical input data conversion and settlement of problems - problems at the stage of making words and kind words. But the increase in the detection of about 4190 entries (words) with a percentage of 6.36% for the detection process encompassing not a compound word and making meaning of the word.

3.3 Suggestions

Suggestions for the next study is the improvement or renewal of encoding in order to detect words and kinds of words that include the meaning of words and word label that serves to determine the use of the word. Further development is done by including a compound word and meaning of words. By adding the meaning of words in a dictionary type the word, the better use, in addition, the meaning of the word can be used to study other NLP requiring the meaning of words such as WordNet development and Word Sense disambiguation.

REFERENCES

[1] C G Ceppy Efraim Bolly, *PEMBANGUNAN KAMUS JENIS KATA SEBAGAI SUMBER DAYA NLP BAHASA INDONESIA*, vol. I, Agustus 2016.

[2] Harimurti, Kridalaksana, *Kelas Kata Dalam Bahasa Indonesia*, 2nd ed. Jakarta: PT Gramedia Jakarta, 1994.