Development of Indonesian Part of Speech Dictionary with Compund Sentence

Yoppy Yansyah
Teknik Informatika
Universitas Komputer Indonesia
Bandung, Indonesia
yoppy787@gmail.com

Ken Kinanti P
Teknik Informatika
Universitas Komputer Indonesia
Bandung, Indonesia
guinev_gs@yahoo.co.nz

In research on "Development of Type Password As a Resource Dictionary Indonesian NLP", resource produced as many as 38 870 entries,consisting of basic words and derivatives, the immature form of the compound as an increase in the resources , Resource compound word is needed for research on NLP as tokenisasi word on POS Tag and question answering. Based on the resource requirements then do kind word dictionary development incorporating compound words that can be used in the development in the field of NLP.

In previous studies the data input KBBI .txt format, so it can not detect entries in bold. Therefore do format conversion process of data input from kbbi.pdf be kbbi.doc and become kbbi.html and done with 3 processes, namely the 5 stages smoothing process input data consists of the elimination karater html entities, removal of labels except label kind words, the elimination of a number meaning, equation symbols and unification compound words separate words. In the process of separating the words do three stages consist of, the separation of compound words, decision words and types of words, and the separation of said base with said derivatives. In the process of detecting a compound word made four stages consist of the elimination of empty rows, filling words and types of words to compound words, cleaning and categorizing.

The end result of the development of a dictionary type of words that include compound words is to produce as many as 51 147 Dictionary kinds of words that include compound words are already available and can be used to study NLP (Natural Language Processing).

Keywords: Dictionary classs word, dictionary online, resource NLP, a compound word.

I. INTRODUCTION

Dictionary is a description of the vocabulary of a language. Dictionary definitions explain how it was to work together to form sentences. The word is a collection of some of the letters that have specific meanings. In KBBI (Dictionary of Indonesian) word is an element of spoken or written language that embodies a feeling and thoughts that can be used in the language. [1] The Indonesian had some various words which are classified by shape, including the base word, the word derivative, a compound word, and the word reset. This compound is a combination of two or more words that form new meanings. For example, if the word green table combined will produce a compound that has a new meaning. [2] Indonesian dictionary since the publication of the first edition in 1988 contains 62,000 entries, the fourth edition in 2008 has the addition of up to 90 049 entries consisting of 41 250 entries and 24 607 principal berimbuhan words, combinations of words as sublema 23 536.

The development of not only the current dictionary in book form, but the dictionary can be accessed digitally. Among others; kbbi.web, kamusbahasaindonesia.org, kateglo.com and others. But now the development of digital dictionaries do not provide resources (resources) and user access rights, so that users can only use alone can not change or add. Resource kind word in the dictionary that include a compound word is needed for the development of the dictionary itself, nor be used for research on NLP (Natural Language Processing) as tokenisasi word on POS Tag and question answering.

In research on "Development of Type Password As a Resource Dictionary Indonesian NLP" [3], resource produced as much as 38 870 entries, consisting of basic words and derivatives, the immature form of the compound word as Extra resources.

Based on the background of the problem that has been presented, the author has a proposal for the development of previous research by adding this type of compound words as a resource (resource) NLP.

II. DATA INPUT

In previous peneletian input data with .txt format can not detect entries in bold, so do the format conversion of the input data is converted into kbbi.doc kbbi.pdf use tools Nitro11. Once converted into kbbi.doc separation data perabjad or one by one from A to Z. The last process, carried out data format conversion from .doc to .html format using the tools html cleaner. Because the input data with the .html format can detect entries in bold.

Table 1 Sample Data Input

1abid ark a kekal; abadi; tidak berkesu- dahan 2abid n IsI

1 orang yg beribadah; 2 ber- ibadah; 3 taat kpd Tuhan; saleh
abilah ark n cacar; ketumbuhan;
-- peringgi penyakit raja singa; sifilis
abece /abécé/ n abc (alfabet Latin)
akasia n pohon atau perdu yg tumbuh di daerah tropis dng rangkaian bunga ku-
ning atau putih;
-- kuning tanaman hias dng bunga ber- warna kuning pd bulir yg bergantung, Acecia auriculiformia
asap n gas yg tampak keluar dr sesuatu yg panas, mendidih, atau dibakar;
mengasap v 1 menjadi asap; 2 memasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap;
~ daging memanggang daging; ~ pakai-
an mengharumkan pakaian dng asap ra- tus dsb; ~ nyamuk mengusir nyamuk dng asap;
baca, membaca v melihat serta memahami isi dr apa yg tertulis (dng melisankan atau hanya dl hati): jangan diganggu, ia sedang ~ buku
mengeja atau melafalkan apa yg ter-tulis;
~ batin membaca dl hati;
~ bibir mencoba mengerti pembicaraan lewat gerak bibir
kolaret /kolart/ n garis, batas, atau rantai yg mengelilingi leher
mahabesar a tidak ada yg melebihi kebe- sarannya; besar yg tidak ada batasnya

III. METHOD AND IMPLEMENTATION

The method used is the method of role-based, process steps performed by three processes, namely the 5 stages of the process of smoothing the input data consists of the elimination karater html entities, removal of labels unless the label kind words, the elimination of a number of meanings, equation symbols of said compound and the unification of separate words. In the process of separating the words do three stages consist of, the separation of compound words, decision words and types of words, and the separation of said base with said derivatives. In the process of detecting a compound word made four stages consist of the elimination of empty rows, filling words and types of words to compound words, cleaning and categorizing.

A. Tidying Data Input

1. Character deletion Html Entity

The removal process karater html entities such as & nbsp; which serves to space mission to help in the process of separating the compound word. because if it is not done it will be considered as a symbol of the end marker (;) will terdeksi

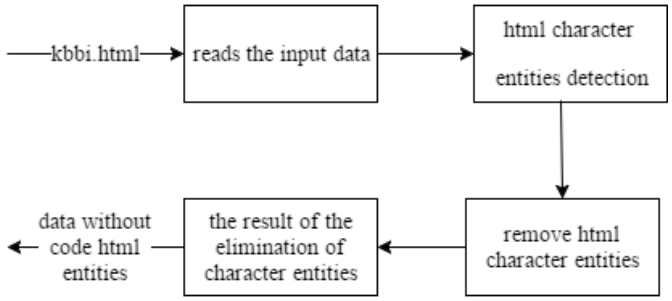


Figure 2 Character deletion Html Entity

Example.

Table 2 Character deletion Html Entity

Before
mengasap v 1 menjadi asap; 2 memasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap; ~ daging memanggang daging; ~ pakai- an mengharumkan pakaian dng asap ra- tus dsb;   ~ nyamuk mengusir   nyamuk dng   asap;
After
mengasap v 1 menjadi asap; 2 memasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap; ~ daging memanggang daging; ~ pakai- an mengharumkan pakaian dng asap ra- tus dsb; ~ nyamuk mengusir nyamuk dng asap;

2. Removal of Labels Except Class Word

The process of removal of labels such words ark, Ar, Filol because it is not necessary to type the word the dictionary results include compound words.

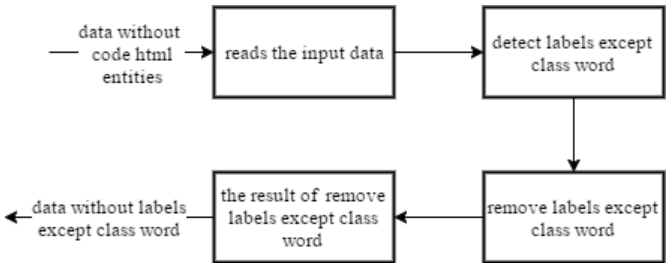


Figure 3 Block Diagram Removal Labels Except Class Word

Example.

Table 3 Except Removal Labels Except Class Word

Before
1abid ark akekal; abadi; tidak berkesu- dahan 2abid n Isl 1 orang yg beribadah; 2 ber- ibadah; 3 taat kpd Tuhan; saleh
After
1abid a kekal; abadi; tidak berkesu- dahan 2abid n 1 orang yg beribadah; 2 ber- ibadah; 3 taat kpd Tuhan; saleh

3. Elimination Number Meaning

The removal process aims to eliminate the number of numbers [0-9] because of the meaning of numbers not required for the results of a dictionary type of compound words.

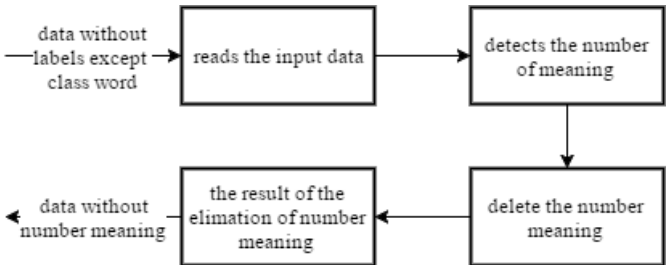


Image 4 Elimination Number Meaning

Example.

Tabel 4. Elimination Number Meaning

Before
1 abid a kekal; abadi; tidak berkesu- dahan 2 abid n 1 orang yg beribadah; 2 ber- ibadah; 3 taat kpd Tuhan; saleh
After
abid a kekal; abadi; tidak berkesu- dahan abid n orang yg beribadah; ber- ibadah; taat kpd Tuhan; saleh

4. Compound Words Symbol Match

The purpose of the equalization process is a compound word symbol change symbol tilde (~) to strip two (-) to change the circuit symbol is a compound word symbol (~) and strip the two (-) are not thick bold dash symbol aimed compound words becomes uniform.

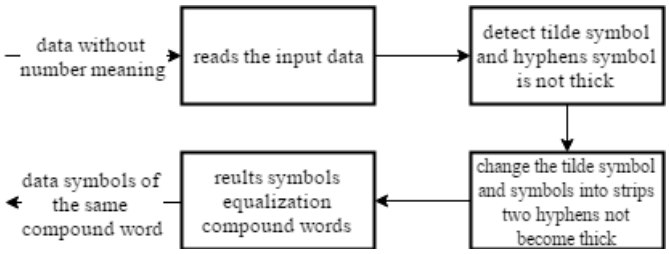


Figure 5 Block Diagram Symbols Match Compound Words

Examples.

Table 5 Compound Words Symbol Match

Before
akasia n pohon atau perdu yg tumbuh di daerah tropis dng rangkaian bunga ku- ning atau putih; -- kuning tanaman hias dng bunga ber- warna kuning pd bulir yg bergantung, Acecia auriculiformia mengasap v menjadi asap; memasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap; ~ daging memanggang daging; ~ pakai- an mengharumkan pakaian dng asap ra- tus dsb; ~ nyamuk mengusir nyamuk dng asap;
After
akasia n pohon atau perdu yg tumbuh di daerah tropis dng rangkaian bunga ku- ning atau putih; -- kuning tanaman hias dng bunga ber- warna kuning pd bulir yg bergantung, Acecia auriculiformia mengasap v menjadi asap; memasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap; ~ daging memanggang daging; ~ pakai- an mengharumkan pakaian

dng asap ra- tus dsb; ~</nyamuk> nyamuk mengusir nyamuk dng asap;

5. The word unification Separated

The purpose of the process of unification of separate words are separate words into whole words.

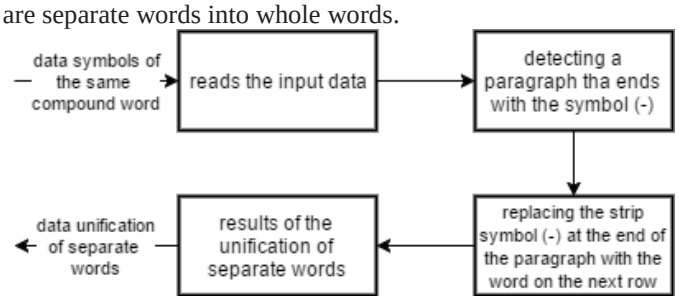


Figure 6 Block Diagram Unification Words Separated

Example.

Table 6 Unification Words Separated

Before
mengasap v menjadi asap; memasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap;   ~daging memanggang daging; ~pakai-   an mengharumkan pakaian dng asap ra- tus dsb; ~nyamuk mengusir nyamuk dng asap;
After
mengasap v menjadi asap; memasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap;   ~daging memanggang daging;~pakaianmengharumkan pakaian dng asap ra- tus dsb; ~nyamuk mengusir nyamuk dng asap;

B. Separating Word

The process of separating the base word by word turunanya and separating the compound word made with three phases. 1. Separation of Compound Words

The process of separation of the compound aimed compound words in a paragraph into paragraphs that have one compound word.

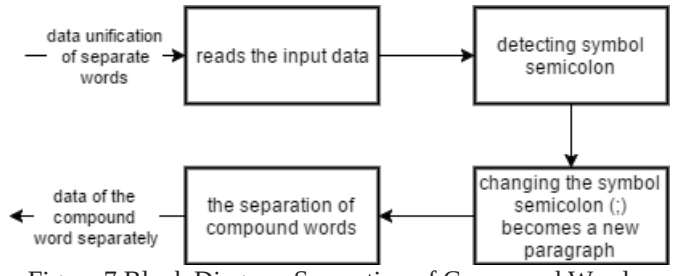


Figure 7 Block Diagram Separation of Compound Words

Example

Table 7 Separation of Compound Words

Before
mengasap v menjadi asap; memasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap;   ~daging memanggang daging; ~pakaianmengharumkan pakaian dng asap ra- tus dsb; ~nyamuk mengusir nyamuk dng asap;
After
mengasap v menjadi asap   memasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap   ~daging memanggang daging   ~pakaianmengharumkan pakaian dng asap ra- tus dsb   ~nyamuk mengusir nyamuk dng asap

2. Take Word and Class Word

The purpose of the process of making words and types of words is taking his words and the type of child tag in paragraph tags.

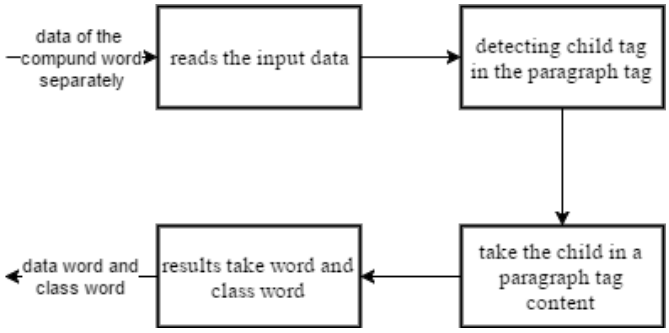


Figure 8. Block Diagram Take Word and Class Word

Example

Table 8 Take Word and Class Word

Before
mengasap v menjadi asap   memasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap   --daging memanggang daging   --pakaian mengharumkan pakaian dng asap ra- tus dsb   --nyamuk mengusir nyamuk dng asap

After
mengasap v --daging --pakaian --nyamuk

3. Separation of Root Words with Words Derivatives

The separation process root word with the word derivative aims to row more than one word and type the word in a row in one word and type the word in the line.

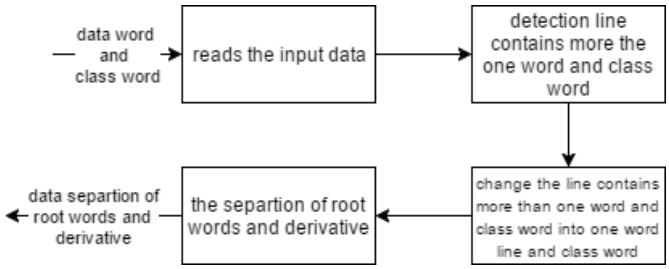


Figure 9 Block Diagram Separation of Root Words with Words Derivatives

Example

Table 9 Separation of Root Words with Words Derivatives

Before
baca, membaca v --batin --bibir
After
baca, v membaca v --batin --bibir

C. Detecting Compound Words

The process of completing a compound word made to fill in the word and a kind word for compound words. Done in 3 stages:

1. Eliminate Empty Rows

The removal process aims to remove blank lines unallocated line sentence.

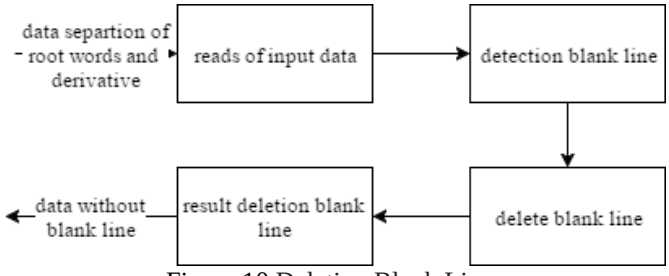


Figure 10 Deletion Blank Line

Example.

Table 10 Deletion Blank Line

Before
abid a abid n abilah n --peringgi
After
abid a abid n abilah n --peringgi

2. Charging Word and Class Word For Compound

The purpose of the charging process and the type of the compound word is replace the two strip symbol (-) into a word on the previous line and fill types main word on the previous line in order to be a compound word.

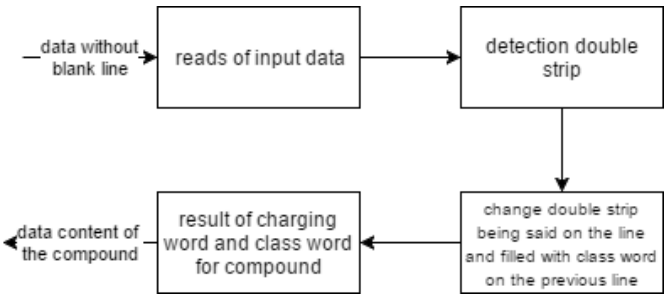


Figure 11 Block Diagram Charging word and Class Word for Compound

Example.

Table 11 Charging word Class Word for Compound

Before
baca, v membaca v --batin --bibir
After
baca, v membaca v membaca batin v membaca bibir v

3. Cleaning

The purpose of the cleaning process is to remove all symbols except the strip. This process is done in the process of completing a compound word because if it is done during the process of smoothing the data input pins html tags will be deleted.

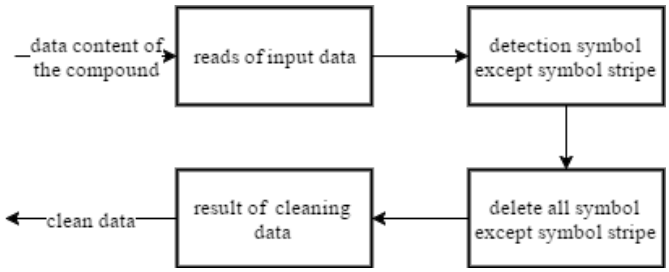


Figure 12 Block Diagram Cleaning

Example

Table 12 Cleaning

Before
baca, v membaca v membaca batin v membaca bibir v koloret /n
After
baca v membaca v membaca batin v membaca bibir v koloret n

4. Categorizing

Intent is to determine the categorization of a compound word and not a compound word with tada description for compound words coupled with the numbers one and not a compound word coupled with zeros.

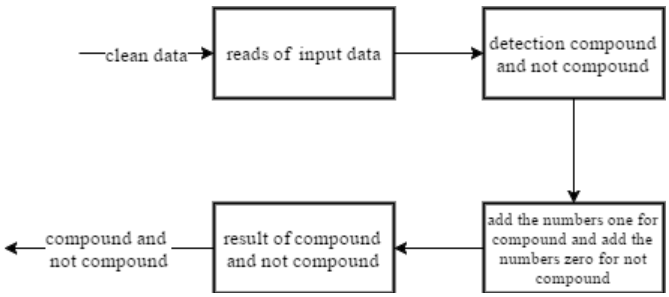


Figure 13 Blok Diagram Categorizing

Example

Table 13 Categorizing

Before
abid a abid n abilah n abilah peringgi n abece n akasia n akasia kuning n asap n mengasap v mengasap daging v mengasap pakaian v mengasap nyamuk v baca v membaca v membaca batin v membaca bibir v koloret n mahabesar a
After
abid a 0 abid n 0 abilah n 0 abilah peringgi n 1 abece n 0 akasia n 0 akasia kuning n 1 asap n 0 mengasap v 0 mengasap daging v 1 mengasap pakaian v 1 mengasap nyamuk v 1 baca v 0 membaca v 0 membaca batin v 1 membaca bibir v 1 koloret n 0 mahabesar a 0

IV. TEST RESULT

Results retrieval compound word and not a compound word consisting of 51147 entries, is described in the table below.

Tabel 14 Result

	Root Word	Derivative	Total
		Compound Word	
Previous reserch	38.870	0	38.870
Current research	40439	10609	51147

In a previous study entry produced as many as 38870 entries. Thus, in this study the number of entries increased by 12277 entries from 38870 into 51147 entries a compound word consisting of 10609 entries and is not a compound word (basic words and words berimbuhan) 40439 entries. Words that are not detected is a word that has no words class, as shown below.

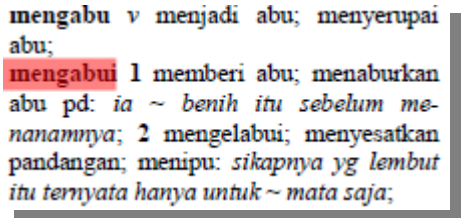

mengabu *v* menjadi abu; menyerupai abu;
mengabui **1** memberi abu; menaburkan abu pd: *ia ~ benih itu sebelum menanamnya*; **2** mengelabui; menyesatkan pandangan; menipu: *sikapnya yg lembut itu ternyata hanya untuk ~ mata saja*;

Figure 14. word that are not detected

V. CONCLUSIONS AND RECOMMENDATIONS

Conclusion The development of a dictionary type of words that include compound words are already available and can be used to study NLP (natural language processing). Of the detection is done, the number of entries generated as many as 51 147 entries consisting of 10 609 and 40 439 entries compound word entry instead of a compound word. This is due to the phase-making process and kind words words are measures that have not detection.

Suggestions for next Penetian is the improvement or renewal measures in order to detect words and kinds of words well. In step categorization of a compound that does not have can not be detected as compound words such enormous and semifinal so the word belongs to the category instead of the compound word. Some words are not detected due to the derivative words that do not have a kind word and therefore to the word derivative that does not have that kind of word, filled with kind words in your base.

REFERENCES

[1] Adi Budiwiyanto, "Korpus dalam Penyusunan Kamus," [Online]. Available: http:// http://badanbahasa.kemdikbud.go.id/. [Diakses 22 Oktober 2016].

[2] Ernawati Waridah, "Ejaan Yang Disempurnakan dan Seputar Kebahasa-Indonesiaan", Bandung : Ruang Kata, 2016

[3] CG Ceppy Efraim Bolly, "Pembangunnan Kamus Jenis Kata Sebagai Sumber Daya NLP Bahasa Indonesia", Bandung, 2016