Developing an Online Sources Of Synonym Ditionary Resources

Januar Saputra
Informatics Engineering
Indonesia Computer University
Bandung, Indonesia
januarsaputra1591@gmail.com

Ken Kinanti Purnamasari
Informatics Engineering
Indonesia Computer University
Bandung, Indonesia
guinev_gs@yahoo.co.nz

A synonym is a word that has different forms but have the sense or the sense of the same or similar. The existence of a resource dictionary of synonyms is currently still in the physical form or the book and not in the form of a database which is accessible for digital or opensource. Plus yet digital dictionaries of synonyms of his output results can be downloaded for free and legal, in other words the user just wear and does not have the permissions to change or add. The existence of a resource dictionary for developers is required, because in addition to help in the process of developing an online dictionary is itself a resource dictionary digital can also be utilized in research related to NLP. Synonym dictionary resources sorely needed as input data in which meaning of the word in the process of disambiguation, analyst sentiment and WordNet.

In this research the source inputs are moderated is English thesaurus dictionary format .txt. Detection of phase jeniskata and synonyms consists of five steps i.e. combining the separate words, normalization of the noise, likens the compound symbol, separate the lines of lema and sublema, ploting and jeniskata and synonym Word format followed generate so kind words and synonyms in the next are stored into the database.

The end result of the construction dictionary of synonyms the synonym dictionary databases are in three formats and online as the dictionary implementation. Pembangunansumber synonym dictionary resources generate 33,379 lema with percentage of 49,984%.

Keywords: Dictionaries, dictionary of synonyms online, NLP resources.

I. INTRODUCTION

Synonym is a word that has a different shape, but has a meaning or definition of the same or similar. Synonym itself can also be referred to as synonyms is a class of words in the language units by category forms the function and meaning in the makeup of the grammatical system [1]. On the use of words such as synonyms have categories more general words, the intensity of meaning, emotive meanings, general and technical, as well as differences in the level of kebakuan

The existence of resources synonym dictionary currently still in physical form or in the form of a book and have a database that can be accessed digitally or opensource. Plus there are no synonyms dictionary its digital output can be downloaded for free and legally, in other words, users simply wear and does not have access rights to change or add. The existence of the resources required dictionaries for developers, because in addition to helping in the development process itself dictionary online digital resource dictionary can also be used in research related to NLP.Sumber synonym dictionary power is needed as input data in the classification of the meaning of words in the process of disambiguation, Analyst sentiment and WordNet.

Based on the needs of resource dictionary that is open source, especially Indonesian, development as a resource synonym dictionary NLP is expected to be one of the resources that can be used in the development of technology in the field of language. Development synonym dictionaries to do with the approach of the field of NLP. The purpose of this research is the implementation of the development of digital synonym dictionary as a resource NLP Indonesian. The objective is to provide the resources for digital synonym dictionary Indonesian NLP research is open source with some sql formats like .txt, .csv, and .json.

II. RESEARCH CONTENTS

data used for the construction of Indonesian synonyms dictionary using a thesaurus dictionary Indonesian agency .txt format.

- *Tesaurus*

Thesaurus is a dictionary Indonesian serving as many words and groups of words and synonyms or synonym said in Indonesian Language.

In any language, there is no similarity absolute meaning. Therefore, synonyms in the Thesaurus Indonesia served by considering the differences that exist based on regional dialects [2].

- *Sinonim*

Synonyms comes from the Greek syn 'with' and ononom 'name'. Although synonyms have the same meaning, but the meaning of the equation it is only a semblance or less the same. So even though the meaning is the same, but still show the differences, especially if associated with the use of the word in a sentence. The similarity of meaning in the word is synonymous not an absolute one hundred percent, but still there is a difference. The difference of meaning can be seen by noting two things: the basic meaning with extra meaning and value of taste [1]

2.1 Identification of problems

Identification of the problem according to the background of the above is the lack of resources digital synonym dictionary Indonesian is open source.

2.2 Purpose and objectives

The purpose of this research is to build digital synonym dictionary Indonesian. The objective is to provide digital resources synonym dictionary Indonesian which is open source.

2.3 Scope of problem

Formulated so that the problems can be more focused, so in this study is limited in terms of the following.

- Data is processed in building a digital dictionary of synonyms is a set of words taken from the Thesaurus Dictionary Indonesian Language Center with .txt format.
- The process of building a digital dictionary of synonyms established only process words are synonyms and derivatives berimbuhan
- Output of system development is synonymous with the online dictionaries that can be downloaded with .sql format .txt, .csv, and .json.

III. METHOD AND IMPLEMENTATION

This type of research is descriptive research that studies the problems that exist and working procedures that apply. The purpose of the study is to create a description of your descriptive, picture or painting in a systematic, factual and accurate information on the facts, nature and the relationship between the phenomenon investigated. In this rulesbased method the researchers gathered the data, arranged in a systematic, factual and accurate. Rulesbased This research method has two stages of data collection and software development. Researchers use this method because it fits with the data and the analysis results from previous research. Development synonym dictionary to be built based on the facts available at the time of data collection and analysis.

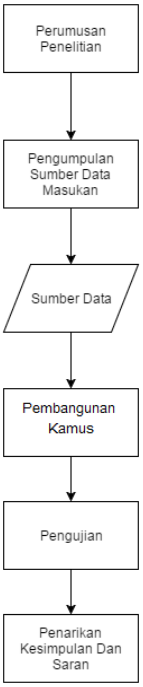


Fig. 1 Scheme Research Methods

3.1 Analysis Solution

In other words, at this stage aims to determine the mechanism of action of a process or procedure that will be built in the implementation of data collection techniques such as resource dictionary of synonyms.

In the analysis of solutions, systems development stages synonym dictionary to be run has the retrieval stage synonyms and kind words. Explanation of each stage can be seen below below

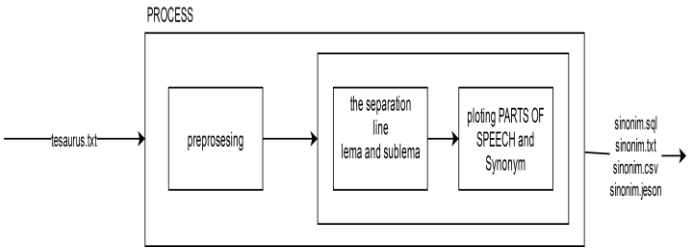


Fig. 2 Overview Main Stage System

In figure 2 System Overview Main Stage will be explained as follows.

- In the early stages the user to enter a text file which is Indonesian Dictionary Thesaurus .txt format into the system
- In the preprocessing stage of the process is carried out as a supporter perapihan dictionary next process
- Furthermore, the detection process jeniskata and synonyms as the final stage which is processed by taking words, kind words and synonyms as an example: cue n cues, cue the N direction, cue n code.

- The last stage is done is the conversion of the database format into several formats including sql .txt, .csv, .dan .json.

3.2 The analysis Data Input

Analysis of the data input to the development system resources synonym dictionary that describes the process of data input of text file data is received and will be processed by the system. Data input in the development of the resources obtained from the synonym dictionary Thesaurus Dictionary Indonesian with .txt format.

3.3 Process analysis

Systematic approach to identifying problems, opportunities, and objectives in designing a system. Analysis of the detection process performed is synonymous words as the main process and data conversion.

3.4 Decision On Synonyms Dictionary Words

Stages decision synonyms performed to obtain the structure of a sentence consisting of words, kind words and synonyms (eg cue n cue). Data input is processed come from Indonesian Dictionary Thesaurus .txt format. Examples of input data are taken from Indonesian Dictionary Thesaurus to be processed is as follows Stages decision synonyms performed to obtain the structure of a sentence consisting of words, kind words and synonyms (eg cue n cue). Data input is processed come from Indonesian Dictionary Thesaurus .txt format. Examples of input data are taken from Indonesian Dictionary Thesaurus to be processed is as follows

aba-aba n arahan, instruksi, isyarat, kode, komando, pe- rintah, petunjuk, seruan, su- ruhan, tanda, titah

abad n era, kala, kurun, masa, periode, sepuluh dasawarsa, sepuluh de- kade, seratus tahun, zaman;

seabad n seera, sekurun, sezaman; berabad-abad adv beratus-ratus tahun, berkurun-kurun, bertahun-tahun, berza- man-zaman

abadi a awet, baka, daim, infinit, kekal, lang- geng, langsung, lestari, qadim, samad, te- tap, wujud; ant sementara

Data input in table Data input is a thesaurus dictionary Indonesian in .txt format the source data to be included in decision synonyms. Of the input data, the system is built to take the word, kind and synonyms eg n cue cue to serve as the next process. Phase Formation carried word format consisting of 10 separate compounding the stages ie, cleanup noise, equalization compound symbol, the separation line and sublema entry and placement of kind words and synonyms.

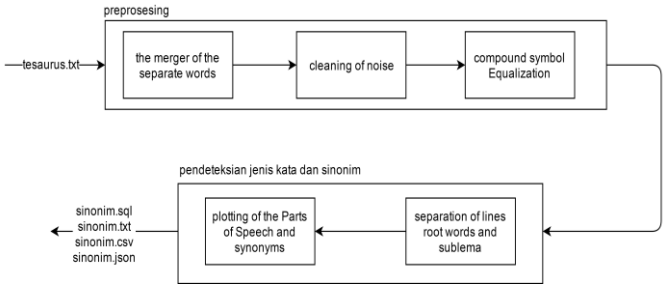


Fig. 3 Intake scheme synonyms

The output of the decision expected synonyms is like in the picture below :

abad n periode

abad n sepuluh dasawarsa

abad n sepuluh dekade

abad n seratus tahun

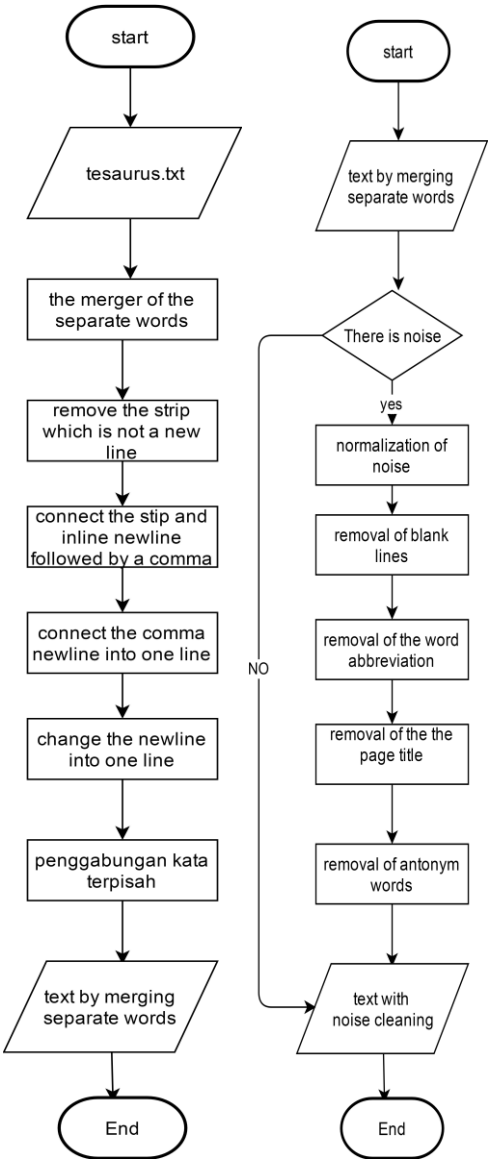
abad n zaman

abadi a awet

abadi a baka

abadi a daim

abadi a infinit



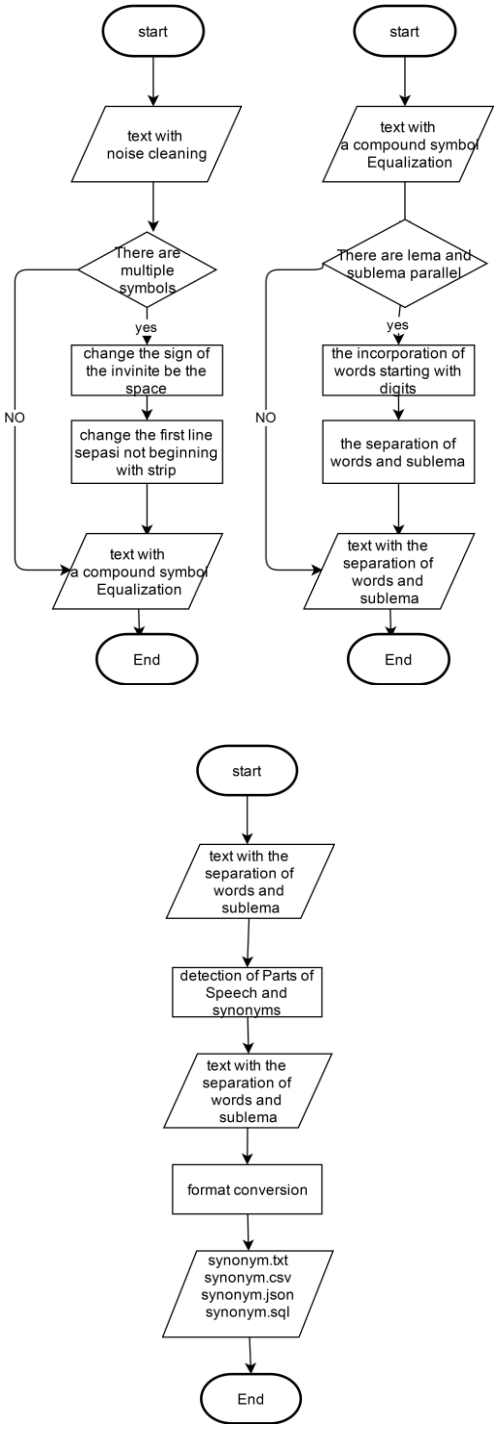


Fig. 4 Flowchart detection synonyms

3.5 format conversion

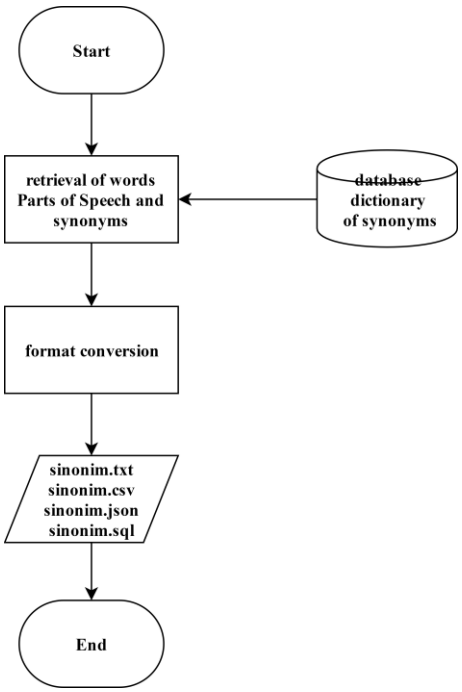


Fig. 5 *Flowchart* format conversion

IV. TEST RESULT

In testing the accuracy will be calculated the value of the accuracy of the number of searches performed entry system based on input data. In its examination the number of words and a kind word obtained will be compared with the number of entries in the dictionary thesaurus Indonesian. Data obtained word entry system is 33 379, while the number of entries in the thesaurus Discuss Indonesia is 66 779 entries.

To obtain a percentage of the captured words and kinds of words on a system that is built, then it was calculated the number of entries generated by the system divided by the number of entries in the thesaurus Indonesian. The next quotient value multiplied by 100% as follows. Jumlah lema Tesaurus Bahasa Indonesia : 66.779 :

- The number of entries generated by the system : 33.379
- Accuracy = $\frac{\text{the number of entries generated by the system}}{\text{Thesaurus entry number Indonesian}} \times 100\%$

  $= \frac{33.379}{66.779} \times 100\%$

  $= 49.984\ \%$

Research conducted generate value accuracy of 49.984 %.

Here is the analysis of test results done within accuracy synonym dictionary development.

In the detection process, said that the condition is said, jeniskata then not followed by synonyms then it can not be stored into the database. Because the data can not be stored influence the outcome of accuracy. Below is an example in the case.

afiat a waras
afiksasi n derivasi
afiksasi n pengimbuhan
afiliasi n
afiliasi n aliansi
afiliasi n asosiasi
afiliasi n hubungan
afiliasi n koneksi

V. CONCLUSIONS AND RECOMMENDATIONS

The system is built, in this case is the construction of a synonym dictionary Indonesian obtain test results development synonym dictionary entries with percentage yield 33 379 49 984%. It can be said there is a lack in the number of entries revenue. Because in the stage of formation precisely in word format to the following five stages of plotting jeniskata and synonyms are not all detected properly.

Suggestions for the next study is the improvement or renewal of encoding in order to detect the format with a better word. Development by including the word antonym as additional resources. The more words are included in the category, the more both in its use.

REFERENCES

[1] Departemen Pendidikan Nasional, Kamus Bahasa Indonesia, IV ed. Jakarta: Pusat Bahasa, 2008.

[2] Eko Endarmoko, Tesaurus Bahasa Indonesia. Jakarta: Gramedia Pustaka Utama, 2006.

[3] Nazir mohammad,Metode Penelitian. Bogor: Penerbit Ghalia Indonesia, 2014.

[4] Rosa A and M Salahudin, *Rekayasa Perangkat Lunak Terstruktur dan Berorientasi Objek*. Bandung: Informatika Bandung, 2014.