BUILDING INDONESIAN WORDNET USING AUTOMATIC TRANSLATION METHOD

Dennis Novriandi[1], Ken Kinanti P. S.Kom, M.T[2]
[1,2] Teknik Informatika – Universitas Komputer Indonesia
Jalan Dipatiukur 114-116 Bandung 40132
E-mail : dennissnov@gmail.com[1], guinev_gs@yahoo.co.nz[2]

ABSTRACT

Wordnet is a network between words that inspired from human lexical memory. It is also described as a place to create an online lexical database to help lexicographer, researcher, or a language activist to search information or the semantic relation of a lemma. Wordnet also can be applied on applications that help human to processing and understanding information like in The Informal Retrieval field, Machine Translation Field, and Natural Language Processing Field.

The development of this Indonesian Wordnet is using an Expand Approach with Automatic translation method. The resource of the data was taken from Princeton Wordnet and using MRD from Cambridge Dictionary.Extraction process was done on the first step so the structured PWN database's data can be obtained. The trasnlation process from English to Indonesian was done using MRD from Cambridge Dictionary in The Second Step.

In the first step, the extracted result data from Unique String is 100 % from 155467 data and Synet is 100% fom 117791 data. On the second step of the translation, 23964 can be obtained from The Unique Strings, and 40774 obtained from The Synset. It happens because of the quality of the MRD itself. There are so many lemma that have no translation and there are many English local word that does not exist in Indonesian.

Keywords : wordnet, Indonesian wordnet, expand approach, Automatic Translation

INTRODUCTION

WordNet research is currently widely researched and developed in different countries, one of Indonesia. Since the Princeton WordNet (PWN) [1] was first discovered by psychology professor George A. Miller in 1985, Over the years a lot of the development of WordNet for other languages by making PWN as a reference for the development. Unlike the other WordNet development, PWN constructed manually with the help of lexicographers, it makes PWN has high accuracy and quality. due to the development of WordNet by hand takes a long time and the costs are high, then use automatic and semi automatic methods as a way to develop WordNet [2]. Research on WordNet Indonesian first ever conducted in Indonesia in 2008 at the University of Indonesia, Indonesian WordNet obstacles yet to be developed back until today. The study resulted in 1203 and 1659 Synset Unique Strings [3]. in 2010 conducted research in the College of Engineering Indonesia using a semi-automatic method that focuses on making Synset without reference to the WordNet of Princeton. Currently PWN have Synset number 117 659 will be used as a resource and a source of data for the development of Indonesian WordNet. Indonesian WordNet development not done again until now. Lack of resources such as lexicographers to Indonesian be a problem for developers Indonesian WordNet, in particular to develop Synset manually.

For the development of the online dictionary itself is still not open source because of the relationship with third parties, so that the limited sources of data become particularly difficult development of WordNet Indonesian dictionary. Usefulness Indonesian WordNet can be used as a reference or source of data for Wordsense Disambiguiton, Information Retrieval, Automatic Text Classification, Automatic Text Summarization, Machine Translation and others. But in the development of Indonesian WordNet constrained from data sources that are still using the data from the outside such as the lack of a complete dictionary of Indonesia for translation. Lack of resource is open source which can be used so it is necessary to research on the development of Indonesian WordNet to provide structured information on WordNet.

Based on the description that has been described above in this study will contribute to develop the earlier research by building applications using the Indonesian WordNet Expand Approach using Automtic Translation. With the help of dwilingual English-Indonesian dictionary for translation entries from PWN to Indonesian. This study is expected to be for a resource in the development of technology and a source of data, especially in particular the development of WordNet into other languages and to the field of NLP.

1. RESEARCH

WordNet using the Expand Development Approach with Atutomatic translation

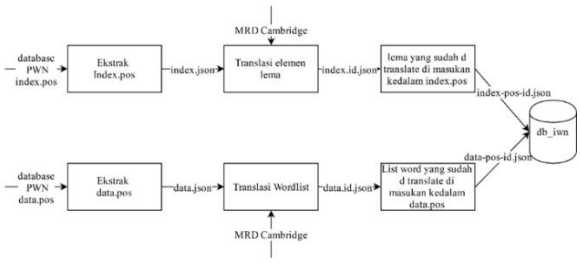
1.1 Analysis of the Problem

For now WordNet Indonesian development has not been done again and the absence of resources that can be used and can be developed by the open source nature. WordNet the need for information is still not

structured due to the development of new dictionaries do so it is necessary to process Indonesian resource. In previous studies Synset generated totaled 1203 and 1659 amounted Unique Strings.

1.2 Solution

The solution provided in this research is using Expand Method Approach to Automatic Translation by translating every entry or words that exist in the database Princeton WordNet (PWN) by way of extracts from Princeton WordNet database and take any semantic relationships that exist in the PWN. The first stage is carried out is the process of cleaning the characters that are not data, such as license agreement, version number and a copyright notice that is at the beginning of each database. Because the database format PWN form and structure data and index data is different for it to be done by two different processes based on the database extract data and index. After a successful extraction stage translational using Machine Readable Dictionary (MRD) from Cambridge Dictionary on any existing entries in the database PWN. Which are stored in the new data. For more details can be seen in Figure 1 Analysis solutions for system development using MRD Indonesian WordNet has a database extract phase PWN, translational entry, and format conversion



Picture 1. Picture Arsitektur Sistem

Overview of the system architecture above is described as follows.

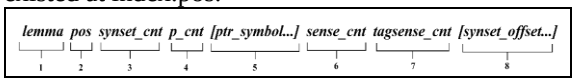
1. In the early stages Users enter PWN index.pos database format into ASCII and extraction system data structures.
2. The translation process is then performed on the first element, namely the lemma. MRD translations from Cambridge Dictionary
3. Lema already translated are then sorted and stored into the database.
4. In this process the input data is a database PWN data.pos ASCII format into the system and then do the extraction and retrieval of data structure, the output format .json
5. At this stage taken wordlist on each line and do the translation from English to Indonesia with the help of MRD.
6. After the translation work then take a data structure that existed at PWN and enter the translation into the structure of the data captured and stored into the database.

1.3 Resources and data input

Analysis of the data input to the database PWN itself has a special format to categorize data by data.pos and index.pos where this post is a class of words that is noun, verb, adjective, and adverb. PWN database format using ASCII encoding so it is easy to be read by humans and machines.

1.3.1 Index.pos

Index file consists of a list of words in the alphabet with each word class database. In every line of the word list have an element named Synset_offset connected with data.pos. each Synset containing the word. Words / entry on index.pos has a lower case format. At index.pos every database consists copyright, version number, license agreement at the beginning of the line. Here is a data structure that existed at index.pos.



Picture 2 Data Input *Index.pos*

The following description of the data structure index.pos..

Tabel 1 Tabel Elemen *Index.pos*

No.	Elemen	Penjelasan
1.	*Lemma*	List words with lowercase ASCII format, spaces are replaced with *underscore*(_).
2.	*Pos*	Pos of tag for each entry, consisting of *n(nount), v(Verb), a(adjective), r(adverb).*
3.	*Synset_cnt*	Number of Synset To determine the amount Synset_offset on index.pos
4.	*P_cnt*	Number of Pointer different for each entry contained in Synset. If not found pointer on sysnet then p_cnt is 0.
5.	*Ptr_symbol*	List Pointers exist in sysnet. Pointer is the semantic relationships that exist in every Synset.
6.	*Sense_cnt*	Number of meaning to every Synset, equal to the number sense sysnet_Cnt
7.	*Tagsense_cnt*	Total the meaning entries that have been in the rank bedasarkan frequent occurrence on semantic relationships.
8.	*Sysnset_offset*	List of offset on data.pos, each

		containing Synset entry / word. each Synset_offset

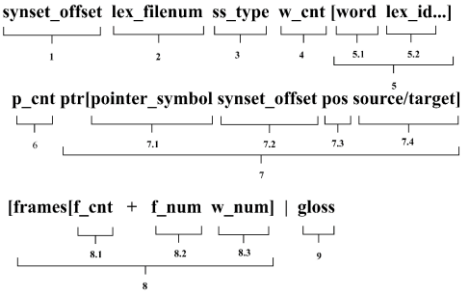
To more clearly following is a sample of the input data index.pos.



Picture 3 Input Data Example *Index.pos*

1.3.2 Data.pos

File data.pos have information that contains members from index.pos Synset. data.pos each file begins with the copyright notice, version number and license agreement. Each row of data has information that has been compiled by lexicographers to Synset. Each data information diawalai by 8 byte offset or address a Synset.



Picture 4 Input Data *Data.pos*

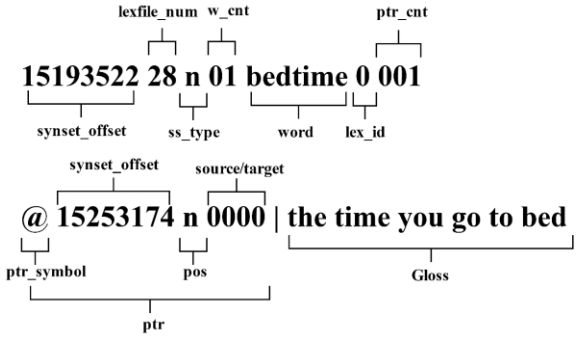
The following description of the data structure data.pos.

Tabel 2 Elemen *Data.pos*

No.	Elemen	Penjelasan
1	*Synset_offset*	8 decimal byte integer containing addresses for information Synset
2	*Lex_filenum*	2-digit decimal integer in accordance with the data of lexicographers containing Synset.
3	*Ss_type*	Pos Tag for each entry, consisting of n (nount), v (Verb), a (adjective), r (adverb).
4	*w_cnt*	2 digit hexadecimal integer that contains the number of words in Synset.
5.1	*Word*	Words with lower case ASCII format. spaces are replaced with an underscore (_). Lodging in data.adj word followed by syntactic marker if the data is specifically
		defined in the data lexicographer.
5.2	*Lex_id*	One hexadecimal digit integer that is the word. Referring to file leksikographer. Lex_id usually starts from number 0 0 terasuk default status and do not have the data lexicographer.
6	*P_cnt*	Three-digit decimal integer that contains the number Pointer to a Synset. P_cnt worth 000 then Synset not have Pointer.
7	*Ptr*	Pointer to a Synset, having the structure: Pointer_symbol offset post source / target
7.1	*Ptr_symbol*	List *Pointer* yang ada pada *sysnet*. *Pointer* adalah relasi semantic yang ada pada setiap *Synset*.
7.2	*Synset_offset*	List Pointers exist in sysnet. Pointer is the semantic relationships that exist in every Synset.
7.3	*Pos*	Pos Tag for each entry, consisting of n (nount), v (Verb), a (adjective), r (adverb).
7.4	*Source/Target*	Source / Target is lexical and semantic Pointer. 4 digit numbering character, the first 2 digits are hexadecimal integer first 2 digits indicate the number of words on Synset. And the last 2 digits indicate the number of words in the target Synset.
8	*Frames*	Frames only in data.verb, has the form F_cnt + f_num w_num
8.1	*F_cnt*	Two-digit decimal integer that menginidikasi number of frames that exist in a row.
8.2	*F_num*	Two digit hexadecimal number an integer that menginidikasi frames
8.3	*W_num*	Two digit hexadecimal integer menginidikasi

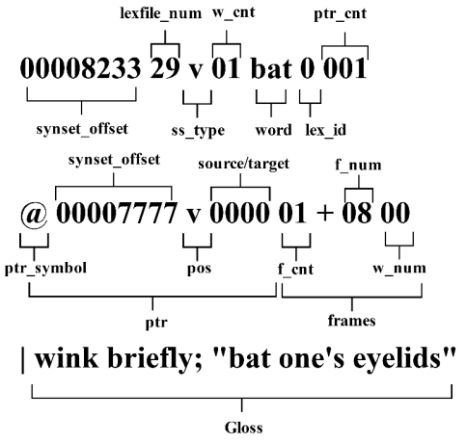
		number of words that contain frames,
9	Gloss	Each Synset has a gloss or definition, separated by the pipe character "|". In gloss elements are definitions and example sentences

Here is an example of input data data.pos....



Picture 5 Example Data Input *Data.pos*

There is little difference to data.verb that is the element frames, Here is the example data.verb for frames.



Picture 6 Example Input Data *Data.verb*

1.4 Process Analysis

The analysis process is conducted is extraction and translation as the main process and data conversion. Because the structure of the data on different index.pos data.pos and then do the extraction process is different to that carried two different processes.

1.4.1 Extraction Index.pos

This section focuses on making the data structure on index.pos to obtain a structure consisting of lemma, post, Synset_cnt, p_cnt [ptr_symbol ...] sense_cnt, tagsense_cnt and Synset_offset [Synset_offset ...]. Data input comes from Princeton WordNet (PWN) format ASCII. PWN itself divides its database into four sections, each of which can be sorted by class said that index.noun, index.verb, index.adverb, and index.adjective Examples of input data are taken from PWN is as follows

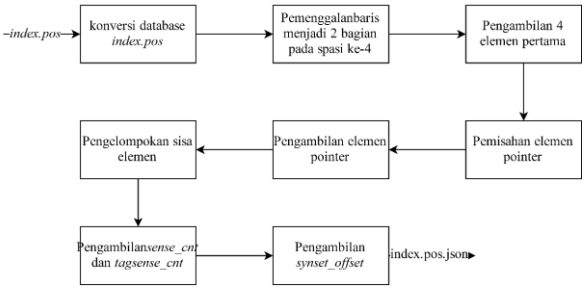
Tabel 3 Data *Index.pos*

```
commerce n 3 6 @ ~ %p = + - 3 1
01092370 08146250 07152330

commando n 2 3 @ ~ ; 2 1 09961754
08421330
```

Tabel di atas dari *database* PWN yang menjadi sumber data yang akan dilakukan ekstrak dan pengambilan struktur data. Tujuanya sistem yang dibuat akan mengambil setiap bagian struktur dan memasukan ke dalam tabel di *database*

Berikut adalah tahapan yang dilakukan dalam ekstraksi database *index.pos*



Picture 7 Extraction Database *Index.pos* from PWN

1.5 Extraction *Data.pos*

After retrieval database index.pos next stage is making data.pos which is containing the data from the index. This stage focuses on making the data structure on data.pos to get the structure of Synset_offset, lex_filenum, ss_type, w_cnt, word, lex_id, [word lex_id ...], p_cnt, [ptr ...], [frames ...] , | gloss,. Data input comes from Princeton WordNet (PWN) format ASCII, the same as its index database PWN divide by post or class of words that data.noun, data.verb, data.adv, and data.adj. sample input data taken from PWN is as follows.

Tabel 4 Example Input Data *Data.pos*

```
08497858 14 n 01 blue 0 002 @
08024893 n 0000 ~ 08411492 n 0000
| any organization or party whose
uniforms or badges are blue; "the
Union army was a vast blue"

08497673 14 n 01 vote 0 000 | a
body of voters who have the same
interests; "he failed to get the
Black vote"
```

7. The input data in the above table is data.pos of PWN database is the source of data to be carried extracts and retrieval of data structures. The goal is the same as the index to take every element. For the data structure is more complex and more difficult than

index.pos for each elemenya not all the same there every row. some differences on any such data.

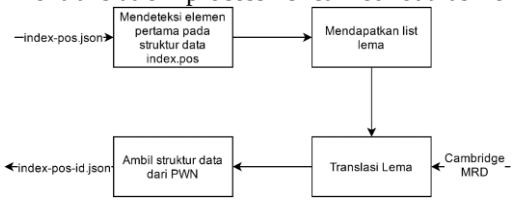
1. [frames ...], the position is before the gloss only on data.verb, and none on the other data.
2. In some cases there are lines that do not have a definition of the word
3. Some of the lines there is no example sentences.
4. Some lines doesn't have pointer

1.6 Translation Index.pos and Data.pos

After the extraction process is carried out every Synset translation on a database that has been extracted. Data input is extracted index.pos and data.pos with .json format. Translation process with translated using Machine readable Dictionary (MRD) from Cambridge Dictionary. Reasons for using MRD Cambridge Dictionary due to the amount of data that can be retrieved from the website. In this process the system will check whether the entries to be translated No terjemahanya or not, if obtained terjemahanya then the first element will be updated and filled with Indonesian translation results, whereas if it is not found, the translation of the entry blank is filled with stuffing. It is useful for validation of how much the number of entries that have been successfully decoded.

1.6.1 Translation Index.pos

In the translation process taken index.pos element is the first element of the data structure, element entries. Unique String entry element is the Princeton WordNet database used to search in Synset members. The translation process is carried out as follows.

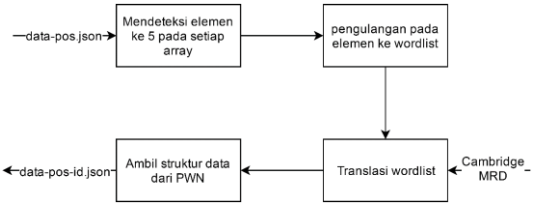


Picture 8 Blok Diagram Proses Translasi *Index.pos*

As we see the structure of the data is taken from the first element of that entry, in the example below there is a list of unique strings taken namely commerce and commando. The system performs translational and then produced the translation to the target language and inserted back into the data structure of the Princeton WordNet.

1.6.2 Translation Data.pos

In the process of translation is taken from the extraction data.pos data.pos, element 5 is all Wordlist. For a list of words Due to the number of words in each line can be more than one, then do repetitions to perform translational afterwards making process from the extraction of data structures. For more details can be seen in the block diagram below.



Picture 9 Block Diagram Translation *Data.pos*

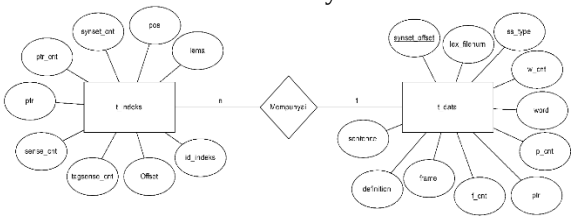
In all five elements in each array be repeated if the word has more than one. Then do the translation of MRD Cambridge Dictionary on the word list, then list the word is entered into the respective data structures for each line.

1.7 Database Analysis

This analysis will be described with ERD and described in the data dictionary ERD.

1.7.1 Entity Relationship Diagram

This analysis will be described with ERD and described in the data dictionary ERD...



Picture 10 *Entity Relationship Diagram* (*ERD*)

1.7.2 Data Dictionary *ERD*

Explanation on Picture Entity relationship diagrams above using the data dictionary ERD, to more clearly can be seen below.

Tabel 5 Kamus Data *ERD*

Entitas	Atribut
t_indeks	*id_indeks, lema,pos,Synset_cnt,ptr_cnt, ptr_sense_cnt, tagsense_cnt, offset.*
t_data	*Synset_offset, lex_filenum, ss_type, w_cnt, word, p_cnt, ptr, f_cnt, frames , definition, sentence.*

1.8 Non Functional Analysis

Non-functional needs analysis was conducted to determine the needs of the specifications for the system that includes hardware and software, this analysis is required to facilitate analyzing a system.

1.8.1 Hardware Analysis

Requirement (hardware) for this development is the minimum specifications to run this system, obstacles to be faced is the process will take a long time. For minimum hardware specs are as follows.

Tabel 6 Requirtment Minimum Hardware

Hardware	Spec Minimun
Prosesor	Intel Pentium IV 3,0 GHz atau AMD Athlon X4
RAM	1Gb

1.8.2 Software Analysis

Analysis software (Software) for the construction of this system are as follows.

Tabel 7 Specification Minimum Software

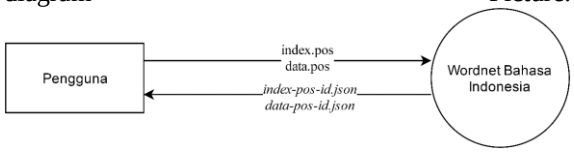
Software	Minimum Recruitment
Sistem Operasi	Windows XP
Web Browser	Google Chrome,mozilla,opera,safari

1.9 Functional Anlaysis

The purpose of this analysis is to analyze the process to be applied in the system to be built using a context diagram, DFD Manufacture and Process Specification

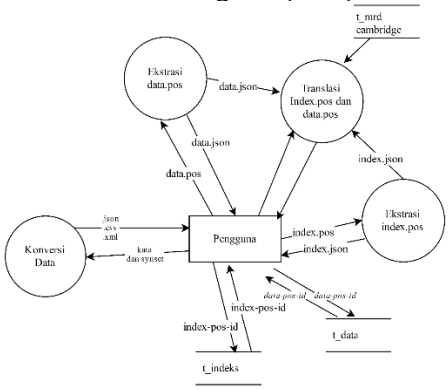
1.9.1 Diagram Context

Diagram context contained in WordNet discussed Indonesia built can be seen in the following context diagram Picture.



Picture 11 Diagram Konteks

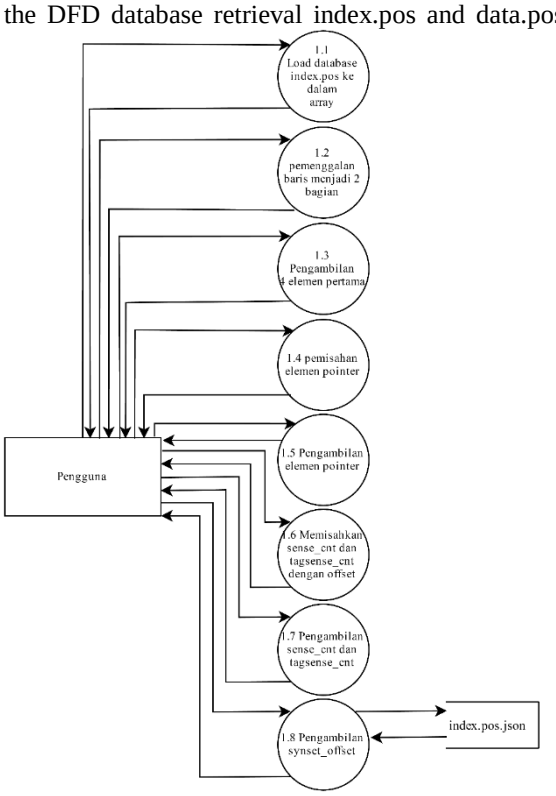
1.9.2 Data Flow Diagram (DFD) Level 1



Picture 12 Data Flow Diagram (DFD) Level 1

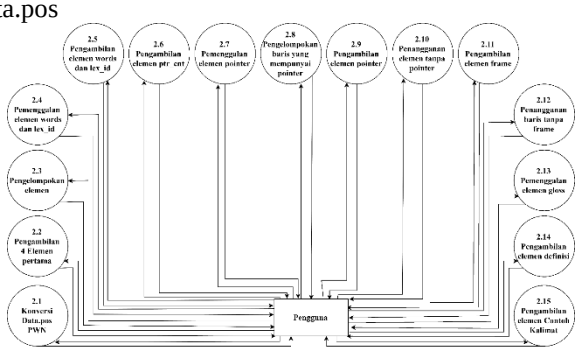
1.9.3 Data Flow Diagram (DFD) Level 2 Index.pos dan Data.pos

Data already entered into decision index.pos and data.pos that diPicturekan in the form of DFD is the flow of incoming data to be generated data. Starting from the extraction of index.pos and data.pos database. The results of this retrieval database is stored in the form of .json which will be processed to the next stage in the form of translational each entry. Data input has its own extension in the form of a post or class of words ASCII format, the data will be generated in the form of .json. The following is from the DFD database retrieval index.pos and data.pos.



Picture 13 DFD Level 2 Proses Ekstraksi *Index.pos*

Here is DFD Level 2 extraction process for data.pos



Picture 14 DFD Level 2 Ekstraksi Data.pos

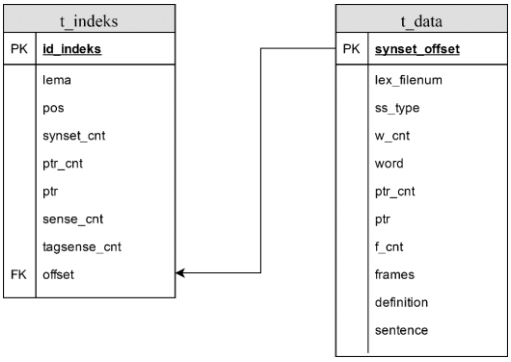
1.9.4 Data Flow Diagram (DFD) Level 2 Translasi Index.pos dan Data.pos

1.10 Database Design

Database design determines the content and arrangement of data required to support a variety of system design, aim to meet the needs of the information contained in the development of Indonesian WordNet.

1.10.1 Relationship scheme

Relation scheme provides relationships between tables in full, including where there is an attribute. It is necessary to store information about objects that are represented in a database. Tables are built for Indonesian WordNet take the data structure of the database Princeton WordNet (PWN). So there are two tables of the index table to list Synset and data tables to data from the Synset. The relation scheme to be built are shown in the following Picture



Picture 15 Relationship Scheme

1.10.2 Structure Table

The tables contained in database design the system to be built is described as follows.

Tabel 8 Structure Tabel t_indeks

Field Name	Data type	Size	Keys	Information
id_indeks	*Integer*	11	PK	*NOT NULL*
lemma	*Varchar*	250		*NOT NULL*
Pos	*Enum*	(‘a’,’n’, ’v’,’r’)		*NOT_NULL*
Synset_cnt	*Integer*	3		*NOT NULL*
Ptr_cnt	*Integer*	3		*NOT_NULL*
Ptr	*Varchar*	250		
Sense_cnt	*Integer*	3		*NOT_NULL*
Tagsense_cnt	*Integer*	3		*NOT_NULL*
offset	*Varchar*	250	FK	

Tabel 9 Structure Tabel t_data

Field Name	Data Type	Size	Keys	Information
Synset_offset	*Varchar*	8	PK	*NOT NULL*
Lex_filenum	*Int*	3		*NOT NULL*
Ss_type	*Enum*	(‘a’,’n’,’v’,’s’,’r’)		*NOT_NULL*
w_cnt	*Int*	3		*NOT_NULL*
Word	*Text*			
P_cnt	*Int*	3		*NOT_NULL*
Ptr	*Text*			
F_cnt	*Int*	3		*NOT_NULL*
Frames	*Text*			
Definition	*Text*			
Sentence	*Text*			

1.11 Accuray testing

Accuracy testing will be calculated on the value of the number of Unique Strings Synset of the results obtained. Synset number that existed at Princeton WordNet is 117 659 of all types of speech. Data taken from index.noun, index.verb, index.adj, index.adv. Here is a table of the results of the extraction and translation of the results of the testing were obtained.

Tabel 10 Acuracy Test

Kategori	Input Data		Extraction		Translation	
	Unique Strings	***Synset***	***Unique Strings***	***Synset***	***Unique Strings***	***Synset***
Noun	117953	82192	117953	82192	11944	20992
Verb	11540	13789	11540	13789	5283	10290
Adjective	21499	18185	21499	18185	4755	7366
Adverb	4475	3625	4475	3625	1982	2126
TOTAL	**155467**	**117791**	**155467**	**117791**	**23964**	**40774**

Accuracy testing will be calculated on the value of the number of Synset of Princeton WordNet database. Synset number that existed at Princeton WordNet is 117 659 of all kinds of speech and the total number of Unique Strings 115467.Dalam pengujianya Synset number who successfully obtained using Expand Approach approach with an automated method is 40 774 and 23 964 Unique Strings Synset

In the first phase extraction results obtained database of PWN accuracy of 100%, all the data and its structure was successfully obtained. In the second phase to get the translation done Synset and Unique Strings into the target language, to obtain percentage sysnset development of the translation by MRD Cambridge Dictionary. Then share the results of calculations performed Synset successful in translation with the number of the PWN synsets which further multiplied by 100%. For its detail is as follows.Jumlah Unique Strings Princeton Wordnet (US-PWN): 155467

Strings Unique number generated by the system (US-IWN): 23964

Accuracy *Unique Strings*

$$= \frac{\text{US-IWN}}{\text{US-PWN}} \times 100\%$$

$$= \frac{23964}{155467} \times 100\%$$

$$= 15.4\ \%$$

Research conducted generate value accuracy of 15.4%

Sedangakan to get Synset accuracy is as follows.

Total Synset Princeton WordNet (S-PWN): 117 791

Synset number generated by the system (S-IWN): 40 774

Accuracy *Synset*

$$= \frac{\text{S-IWN}}{\text{S-PWN}} \times 100\%$$

$$= \frac{40774}{117791} \times 100\%$$

$$= 34.6\ \%$$

Research generate value accuracy of 34.6 %.

1.11.1 Analysis of the Test

Here is an analysis of the accuracy of test results done in the development of Indonesian WordNet.

The data generated is of the category noun, verb, adj and adverb results obtained to Synset amounted to 40 774 of the total 117 791 and produced 23 964 Unique Strings of 155467. In the translation process can be seen in the noun category having a large impact on the percentage of test results just obtained around 10% in this case because the Unique Strings her no word translation of the translation using MRD Cambridge Dictionary because the noun form of the word is a term mostly nouns that are regional in English. Many of the terms Medical, Chemical, and other special terms to the dictionary that can not be translated and compound words in the English language very little translation in Indonesian language.

2. CONCLUSION

On This section describes the conclusions and recommendations of the research conducted.

2.1 Conclusion

The system is constructed in this case is the development of Indonesian WordNet has been successfully created. And the availability of a lexical database is open source. In the first phase of the extracted Unique Strings get as much as 155 467 of the total initial data Synset 155 467 and 117 791 of the total of data as early as 117791. In the second stage in the process of translation obtained Unique Strings Synset as many as 23 964 and 40774. Can be seen on the extraction results obtained extraction entirely in the first stage and the second stage occurs shrinkage results due to the quality of the MRD. Many entries as the word "Batman" that no terjemahanya and many local terms for the English language that do not exist in Indonesian.

2.2 Future Works

Suggestions for future research is adding an element meaning of KBBI and the addition of Unique Strings and Synset enhancement with automatic methods or the construction of semi-automatic Synset using the thesaurus and KBBI.

BIBLIOGRAPHY

[1] G. A. M. (. W. A. L. D. f. English., "WordNet: A Lexical Database for English.," WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41. , 1995.

[2] P. Vossen, EuroWordNet: A Multilingual Database with Lexical Semantic, Publishers, 1998, 179 pp; hardbound,, 1998.

[3] A. A. a. R. M. Desmond Darma Putra, Building indonesian Wordnet, Universitas Indonesia, 2008.

[4] B. M. E. P. Drs. Sumardi Suryabrata, Metodelogi penelitian.

[5] I. Sommerville, Software Engineering (Rekayasa Perangkat), Erlangga, 2011.

[6] D. A. Chaer, Tata Bahasa Praktis Bahasa Indonesia, Jakarta: Rineka Cipta, 2011.

[7] M. Dr. R Kunjana Rahardi, Bahasa Indonesia Perguruan Tinggi, Erlangga, 2015.

[8] R. beckwith, G. A. Miller and R. tengi, Design and Implementation of the WordNet Lexical Database, 1993.

[9] R. A dan M. Salahudin, dalam Rekayasa Perangkat Lunak Terstruktur dan Berorientasi Objek, Bandung, Informatika Bandung, 2014.

[10] c. felabaun, English Verb as semantic net, 1993.

[11] A. Solichin, Pemrograman Web dengan PHP dan MySQL, 2016.

[12] M. N. S. S. B. Nurril Hirfana, Creating the Open Wordnet Bahasa, Nanyang university.