

Stemming Algorithm for Indonesian Signaling Systems (SIBI)

Risky Aswi Ramadani¹, I Ketut Gede Darma Putra^{2*}, Made Sudarma³, and Ida Ayu Dwi Giriantari⁴,
¹Prodi Teknik Informatika, Fakultas Teknik, Universitas Nusantara PGRI Kediri
^{2,3,4} Program Doktor Ilmu Teknik, Universitas Udayana
 * Email: ikgdarmaputra@unud.ac.id

Abstract The Indonesian Language Sign System (SIBI) is a medium used by deaf people to communicate with the wider community. SIBI is Indonesian that is presented by hand. The sentence structure (morphology) of SIBI is the same as the Indonesian language, SIBI also recognizes the prefix, principal and suffix sign word. This research discusses how to make stemming for SIBI, the algorithm used for stemming is Nazief Adiri. Stemming for SIBI separates input sentences, words through the Text Processing process. Then the word in Stemming so that it can be separated into prefix, main, and suffix words. After stemming, the words obtained in indexing can find the right SIBI, by using ID. Good stemming must pay attention to Indonesian Morphology. The number of words used for this stemming trial is 50 words. By using the Confusion Matrix method found 0.94% accuracy, 93% recall, and 84% precision.

Index Terms—Accuracy, Stemming, Text, Morphology, Indonesian Sign Language System (SIBI).

I. INTRODUCTION

SIBI is a medium used by people with hearing impairment to communicate with outside communities, be it the general public or fellow deaf people. The SIBI morphology is the same as the Indonesian language, because SIBI is the Indonesian language which is presented by hand gestures [1]. SIBI consists of 2 scopes, namely principal and additional sign word.

Principal sign word are signs that symbolize a word or concept, these sign word are formed with a variety of performers, places, directions and frequencies. While the additional signals are signals that symbolize the beginning, and the suffix [2] [3]. It can be concluded that the morphology of Indonesian and SIBI are the same.

At this time, especially in the Indonesian language, there are very many studies that discuss stemming, but there is no SIBI to date [4] - [6]. Whereas the need for stemming for SIBI is very much needed especially for the development of SIBI translators [7] [8].

The solution offered by this research is to make a Stemming using the Nazief Adriani Algorithm. Nazief Adriani Algorithm is an algorithm used to search for headwords [9] - [11]. Because the needs of this research are not just the main words. This research will show the main

words, prefix words, and suffix words [12] [13].

The Stemming process itself is used to separate words, becoming principal words, prefix words, and suffix words. For example, to separate the word "read". The word reading consists of the prefix word "mem" and the main word "read". Words obtained from stemming are then indexed and then matched with SIBI images.

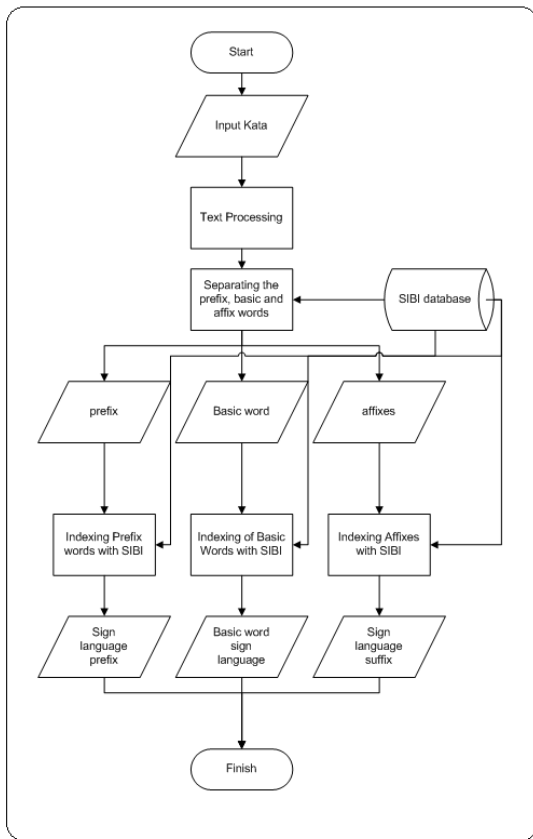
The SIBI image has been examined in 2018, with the title of the article "Database of Sign Language". After the data is indexed, the word data in SIBI is displayed. In this study the presentation of words in SIBI is in the form of images [14] [15].

The main words, prefixes, and suffixes in this study will be displayed in SIBI. The novelty of this research is to create a special SIBI Stemming using Nazief adriani algorithm.

This research is really needed by people with hearing impairment, besides this research is the basis of subsequent studies relating to the translator's tool for the deaf. Therefore the success rate of this study needs to be measured. To measure the level of success will be used Confusion Matrix. By using Confusion Matrix, accuracy, precision and recall values can be found [16] - [18].

II. RESEARCH METHOD

SIBI Stemming has the following stages: the stages that will be passed.



A. Input Words

Word input is the initial stage of the Stemming process, words that are inputted are words that contain a prefix, or words that contain a suffix. Words that get a prefix or suffix usually have a change of meaning. Also in SIBI, there is no word that immediately gets a prefix. For example in this section there is the word "eat" this word consists of the prefix "me(me)", and the main word "(makan)eat" [19]. To get results like that the Stemming process needs to be done.

B. Text Prosesing

Before the sentence is stemming the sentence through the stages of Text Processing [20] [21], the following are the stages of the text processing.

1. Case Folding

Case folding is used to convert text into a standard form so that it is easily understood by computers, for example "Memakan(Eating)" to "memakan(eating)"

2. Tokenizing

Tokenizing is used to separate sentences based on spaces so that sentences become words. The following is an overview of the Tokenizing Process.

3. Filtering

Filtering is used to discard words that are less important, and have no meaning. The condition of the filtering process is that it cannot change meaning. Word data that is discarded contained in a table called a stoplist.

C. Stemming (Separating the prefix, the main, and the suffix)

Actually the stemming function is used to know only basic words. While the word prefix, and the word affix are removed. But in this study the words prefix and suffix are not removed but only separated, but still displayed. Because this stemming is used specifically for translators SIBI [22]. If the word prefix and the suffix are omitted, a meaning will occur.

D. Output Proses Stemming

In this Stemming Process words are divided into 3 parts namely prefix words, basic words, suffix words.

a. The word prefix is the word in front of the base word.

The following are the basic words in SIBI ber-, di-, ke-, me-, pe-, se-, ter-.

b. The main word, which is a word used to describe a word or concept, is formed by various winnings, places, directions, and frequencies as described above.

The word "memakan(ea)t" is a word that comes from the main word "makan(eat)", then the word gets the prefix "me" and the suffix "kan".

E. Proses Indexing


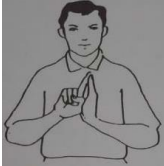

This indexing process is used to change words in the form (text) into words (SIBI). In order to change the word to this text correctly, use Id.

F. Output of prefix sign word, principal sign word, and suffix sign word

SIBI's morphology is the same as Indonesian morphology. Didala SIBI is also known as the prefix sign word (prefix words), principal sign word (principal words), and suffix sign word (suffix words) [23] [24]. The following is a more detailed and more detailed explanation.




a. Isyarat Awalan(kata Awalan).

The prefix gesture is the word in front of the root word. The following are the basic words in SIBI ber-, di-, ke-, me-, pe-, se-, ter-. But the sign word was pronounced with the movement of the hand. The following is a picture of the prefix signs at SIBI.

no	Word	sign language prefix
1.	Ber-	
2.	Di-	
3	Me-	




b. Sign Language Basic words

Signs that symbolize a word or concept Signs are formed with a variety of performers, places, directions, and situations. The following are some of the Basic Sign Language

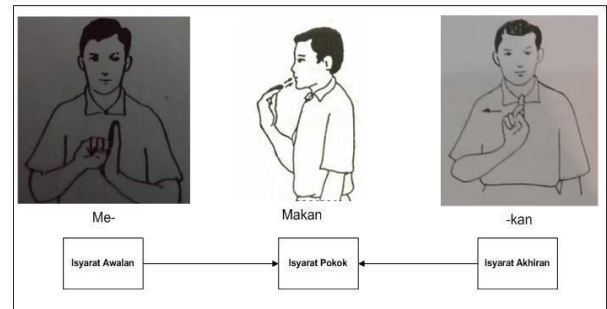
no	Word	Sign Language Basic words
1.	Makan(eat)	
2.	Tidur(sleep)	
3.	Meja	

c. Sign Language suffix (suffix)

Formed sign word are signals that are formed by combining principal signatures with affixed sign word and by combining two or more principal sign word. The following are examples of suffix signaling sentences.

No	Word	Sign Language suffix
1.	-i	
2.	-kan	
3.	-wan	

In the process of word stemming, the structure of the isyarah is divided into 3 parts, namely the prefix sign word, the principal sign word, and the suffix sign word. The following is an illustration of the word division.



III. ANALISA

.Stemming for Indonesian Signaling Systems (SIBI) has a different character, with stemming in general. Stemming generally only shows the main word, does not display the prefix, and the ending word, which is displayed only the main word. Whereas in SIBI, steming must continue to display the prefix, principal, and suffix. The aim is to divide the meaning, which is caused by the lack of an affix or suffix. The following is a table that explains the steming process which transforms a word into a prefix, a main word, and an accent word.

Tabel 1. Word Cutting

No	Word	Affixes word	Basic words	Suffix words
1.	Membaca(Read)	Mem-	Baca(Read)	-
2.	Memakan(eat)	Me-	Makan(eat)	
3.	Biarkan(let it be)	-	Biar(let it be)	-kan
4.	Belajar(Learn)	be	Ajar(teach)	
5.	Ajarkan(Teach)	-	Ajar(teach)	-kan

The first step taken is the input of words (text) for example consuming words, then the words go through the process of stemming. After the prefix, main, and affix words are separate. The words are indexed with SIBI. After the Signing Indexing is displayed based on the prefix sign language, principal sign language, and affix sign language.

Proper testing needs to be done because in the process of stemming accuracy. Precision and recall are very necessary. To fulfill these three aspects, the matrix confusion method is used. The following is the test scenario. Stemming performance test for SIBI, tasked with recognizing words that contain affixes from several words entered so that words are divided into prefixes, principal words, and ending words. To test 100 words that have prefixes and prefixes will be included and words that do not have additive as many as 3340 Results The words identified have a prefix and an affix of 118. After an evaluation it turns out that the number of words that have a prefix is only 92. While 20 other words are words that do not have a prefix, ending or error when recognition of words. In order to be easily understood the scenario is translated in the form of a table. Here is the Confusion Matrix table

Tabel 2. Confusion Matrix

		True Value	
		True	False
Prediction Value	True	92	20
	False	8	3320

The values in table 1 are processed using the confusion matrix method so that the values of precision, recall, and accuracy are found. The following is the calculation.

$$Presisi = \frac{92}{92+20} = \frac{92}{112} = 0,82 \dots \dots \dots (1)$$

$$Recall = \frac{92}{92+8} = \frac{92}{100} = 0,92 \dots \dots \dots (2)$$

$$Akurasi = \frac{92+3320}{92+3320+20+8} = \frac{3412}{3440} = 0,991 \dots (3)$$

$$92+3320+20+8 \quad 3440$$

From the calculation we get a precision of 82% and an accuracy of 99%. Recall of 92%. From the results of the above calculations it can be stated that the system is able to work very well, because it is able to recognize words that have the word affix and suffix. Besides that the system can separate well.

IV. CONCLUSION

The Indonesian language system (SIBI) has the same morphology as the Indonesian language. In SIBI there are also the words prefix, principal and suffix. In some problems stemming is used to find the main word, without displaying the prefix and akiran words. In this study, what is done is different because of the need for this research to find out the syllable of a word that gets the word prefix and suffix. This word separation uses the Stemming method. From the results of stemming analysis can work well, this can be proven by the high number of recall which is 92%.

REFERENCES

[1] Departemen Pendidikan Nasional, "Kamus Sistem Isyarat Bahasa Indonesia," . Jakarta., 2002, pp. xiv.

[2] M. Retno, Meningkatkan Kemampuan Memahami Bacaan Pada Pelajaran Bahasa Indonesia Dengan Media Sistem Isyarat Bahasa Indonesia (Sibi) Siswa Kelas Dasar 2 (D2) Slb-B Yakut Purwokerto Tahun Pelajaran 2008/2009. UNS, Surakarta , 2009, pp. 1-77.

[3] J.L. Luckner, and C. Cooke "A Summary of the Vocabulary Research With Students Who Are Deaf or Hard of Hearing," *American Annals of the Deaf.*, 2010, pp. 36-67.

[4] G. Septian, A. Susanto and G. F. Shidik , *Indonesian news classification based on NaBaNA* . International Seminar on Application for Technology of Information and Communication (iSemantic) , Semarang, 2017, pp. 175-180, doi: 10.1109/ISEMANTIC.2017.8251865.

[5] D. S. Maylawati, W. B. Zulfikar, C. Slamet, M. A. Ramdhani and Y. A. Gerhana, An Improved of Stemming Algorithm for Mining Indonesian Text with Slang on Social Media, International Conference on Cyber and IT Service Management (CITSM), Parapat, Indonesia 2018, pp. 1-6

[6] D. E. Cahyani, L. Merta Tri Utami and H. Setiadi, Clustering of Javanese News in Krama Alus Level with Javanese Stemming, International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 2019, pp. 462-467.

[7] M. N. Baehaqi, M. Irzal and F. H. Indiyah, Morphological Analysis of Speech Translation into Indonesian Sign Language System (SIBI) on Android Platform, International Conference on Advanced Computer Science and Information Systems (ICACSIS), Bali, Indonesia, 2019, pp. 205-210

[8] F. San, Investigation of Translator Training Departments' Course Contents with Regards to Language of Instruction, Social and Behavioral Sciences, 2015, pp. 869-876.

[9] H. Joshi, J. Pareek, R. Patel and K. Chauhan, To stop or not to stop - Experiments on stopword elimination for information retrieval of Gujarati text documents, International Conference on Engineering (NUiCONE), Ahmedabad, 2012, pp. 1-4

[10] J. K. Raulji and J. R. Saini, Generating Stopword List for Sanskrit Language, International Advance Computing Conference (IACC), Hyderabad, 2017, pp. 799-802.

[11] S. R. Manalu, Stop words in review summarization using TextRank, International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Phuket, 2017, pp. 846-849.

[12] A.K. Fahmi, Analisis Kesalahan Gramatikal Teks Terjemah (Indonesia-Arab) Dalam Pendidikan Bahasa Arab, 2016, pp. 105-116

[13] I. Ilmi, Morphological Errors On Arab-Indonesia Translation Text Using Google Translate, Journal Arabic Learning, 2019, pp. 175-185

[14] R. A. Ramadhani, I. K. G. D. Putra, I. M. Sudarma and I. A. D. Giriantari, Database of Indonesian Sign Systems, International Conference on Smart Green Technology in Electrical and Information Systems (ICSGTEIS), Bali, Indonesia, 2018, pp. 225-228

[15] J. L. Garcia-Balboa, M. V. Alba-Fernandez, F. J. Ariza-López and J. Rodriguez-Avi, Homogeneity Test for Confusion Matrices: A Method and an Example, International Geoscience and Remote Sensing Symposium, Valencia, 2018, pp. 1203-1205.

[16] L. Chen and H. L. Tang, Improved computation of beliefs based on confusion matrix for combining multiple classifiers, Electronics Letters, vol. 40, no. 4, pp. 238-239, 19 Feb. 2004

[17] F. J. Ariza-Lopez, J. Rodriguez-Avi and M. V. Alba-Fernandez, Complete Control of an Observed Confusion Matrix, Geoscience and Remote Sensing Symposium, Valencia, 2018, pp. 1222-1225

[18] Y. Xiong, Building text hierarchical structure by using confusion matrix, International Conference on BioMedical Engineering and Informatics, Chongqing, 2012, pp. 1250-1254.

[19] Kamus Bahasa Indonesia, Pusat Bahasa, Jakarta, 2008.

[20] X. Lin, J. Yang and J. Zhao, The text analysis and processing of Thai language text to speech conversion system, International Symposium on Chinese Spoken Language Processing, Singapore, 2014, pp. 436-436.

[21] S. Shi, T. Cheng, S. Xiao and X. Lv, "Text Processing in Video Frames with Complex Background," 2009 International Forum on Information Technology and Applications, Chengdu, 2009, pp. 450-454

[22] R. B. S. Putra and E. Utami, Non-formal affixed word stemming in Indonesian language, International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, 2018, pp. 531-536

[23] K. Khotimah, Analysis Of Indonesian Affixes In English Words Found In Mobile Guide Edition: 54-59, Universitas Diponegoro Semarang, Semarang, 2012..

[24] Anonim, Kamus Sistem Isyarat Bahasa Indonesia, Departemen Pendidikan Nasional, Jakarta, 2001.