JOIN (Jurnal Online Informatika)
Volume 6 No. 2 | December 2021: 218-224
DOI: 10.15575/join.v6i2.812
p-ISSN: 2528-1682
e-ISSN: 2527-9165

Sundanese Stemming using Syllable Pattern

Ade Sutedi[1], Rickard Elsen[2], Muhammad Rikza Nashrulloh[3]
[1,2,3]Department of Computer Science, Institut Teknologi Garut, Indonesia

Article Info

Article history:

Received September 15, 2021
Revised October 18, 2021
Accepted November 09, 2021
Published December 26, 2021

Keywords:

Phonology
Stemming
Sundanese
Syllable

ABSTRACT

Stemming is a technique to return the word derivation to the root or base word. Stemming is widely used for data processing such as searching word indexes, translating, and information retrieval from a document in the database. In general, stemming uses a morphological pattern from a derived word to produce the original word or root word. In the previous research, this technique faced over-stemming and under-stemming problems. In this study, the stemming process will be improved by the syllable pattern (canonical) based on the phonological rule in Sundanese. The stemming result for syllable patterns gets an accuracy of 89% and the execution of the test data resulted in 95% from all the basic words. This simple algorithm has the advantage of being able to adjust the position of the syllable pattern with the word to be stemmed. Due to some data shortage constraints (typo, loan-word, non-deterministic word with syllable pattern), we can improve to increase the accuracy such as adjusting words and adding reference dictionaries. In addition, this algorithm has a drawback that causes the execution to be over-stemming.

Corresponding Author:

Ade Sutedi,
Department of Computer Science,
Institut Teknologi Garut,
Jl. Mayor Syamsu No. 1 Garut, Indonesia
Email: adesutedi@itg.ac.id

1. INTRODUCTION

There are differences in speech and word writing in the Sundanese language in several regions, but the Sundanese Priangan dialect has become the standard for Sundanese society, especially in West Java [1]. However, the data resources and studies such as text mining, machine learning, and natural language processing for Sundanese are limited. It is because the influence of foreign languages and Indonesian as the official language has strengthened [2]. Therefore, this research is driven by the lack of corpus data availability and development activities in software tools used for various applications, especially those related to the Sundanese language. However, recent studies that discuss stemming techniques for regional languages have begun such as Javanese [3][4], Madura [5], Balinese [6][7], and also Sundanese [8-10].

The stemming technique in previous research was developed in morphology form including derivation and inflectional affixes removal for regional language [3-10], Indonesian [11][12], and English [13]. Several stemming methods used in previous research are rule-based [3][10][13], the combination of rule-based with corpus or dictionary [4-12], also rule-based with n-gram [7]. The rule-based technique adopted from Porter's algorithm focuses on the process to remove the suffix to the right of the root word [13]. This method is widely applied to stemming development for Indonesian and later adapted to stemming for regional languages. Another technique was developed which applied to several regional languages stemming using confix stripping [11] and enhanced [4][5] for confix stripping (ECS) method. Although the ECS technique has good stemming results for Indonesian, but not for regional languages.

Meanwhile, stemming for the Sundanese language becomes difficult when it covers allomorph [10] which could be a kind of base word or original word [8]. In addition, the difference of morphological patterns in regional languages has many differences compared with Indonesian or English language such as derived words constructed by the combination of prefixes, infixes, suffixes, or confixes [9][10]. So, the application of the rule-based method becomes ineffective for some cases especially if it depends on the completeness of the

dictionary database [4-12] because there are even some words that have no equivalent in Indonesian [15]. To overcome this problem, the research uses a different rule-based approach by utilizing syllable patterns that form derived words in the Sundanese language. This process is expected to be able to improve several cases that have not been resolved in previous studies.

2. METHOD

The study of the sound of language becomes the basis of writing or grammar. Phonemes in Sundanese can have a variety of pronunciations corresponding to their place in the noun or word, typically not distinguishing meaning [15]. There are twenty five phonemes in Sundanese[15][17], which divide into two categories namely vowel (Swara) and consonant or append letter (Wianjana). For the 7 vowel distribution i.e.: /a/ [a], /i/ [i], /u/ [u], / é / [ɛ], /o/ [o], /eu/ [ö], /e/ [e] and 18 consonants, i.e.: /b/ [b], /c/ [c], /d/ [d], /g/ [g], /h/ [h], /j/ [j], /k/ [k], /l/ [l], /m/ [m], /n/ [n], /ny/ [ñ], /ng/ [ŋ], /p/ [p], /r/ [r], /s/ [s], /t/ [t], /w/ [w], and /y/ [y]. Along with the development of national and foreign languages uses, there are additional consonants, i.e.: /f/ [f], /q/ [q], /v/ [v] , /x/ [x], and /z/ [z] [17].

2.1. Research Flow

In this research, the stemming process using syllable pattern algorithm was developed based on the following research flow depicted in Figure 1.

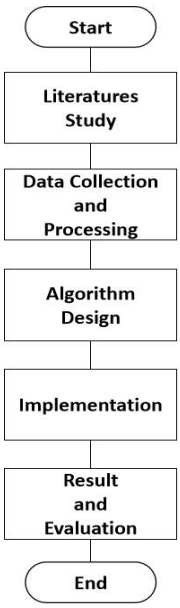

Figure 1. Research Flow for Stemming using Syllable Pattern

2.2. Syllable Pattern

Syllables are the basic word-forming components (morphemes) in the Sundanese language which generally consist of two syllables. Generally, the Sundanese language consists of vowels and consonant arrangements as a system of sound in the form of syllables or canonic. The canonical form of the word in the Sundanese language is as follows [14]:

V - Vowel
KV - Consonant Vowel
VK - Vowel Consonant
KVK - Consonant Vowel Consonant
KKV - Consonant Consonant Vowel
KKVK - Consonant Consonant Vowel Consonant

In addition, due to the influence of the use of national and foreign languages, there are several canonical patterns as follows:

VKK - Vowel Consonant Consonant
KVKK - Vowel Consonant Consonant
KKKV - Consonant Consonant Consonant Vowel
KKVKK - Consonant Consonant Vowel Consonant Consonant
KKKVK - Consonant Consonant Consonant Vowel Consonant

KVKKK - Consonant Vowel Consonant Consonant Consonant

For all Sundanese syllable pattern, the writing rules are as follow [16]:

1. Sundanese words consist of *Ekasuku* (one of the syllables), *Dwisuku* (two of syllables), *Trisuku* (three of syllables), *Catursuku* (four of syllables), and *Pancasuku* (five of syllables).

Table 1. Canonical Form in Sundanese

Ekasuku	*Dwisuku*	*Trisuku*	*Catursuku*	*Pancasuku*
V	V-V	V-V-KV	V-KV-KV-KV	V-KV-KV-KV-KVK
VK	V-KV	V-KV-KV	V-KV-KV-VK	KV-KV-KV-KV-KV
KV	V-KVK	V-KV-KVK	V-KVK-KV-VK	KV-KV-KV-KVK-KV
KVK	V-KKV	VK-KV-KV	V-KV-KV-KVK	KV-KV-KV-KVK-KVK
KKV	V-KKVK	VK-KV-VK	VK-KV-KV-KVK	KVK-KV-KV-KV-KVK
KKVK	VK-VK	VK-KV-KVK	VK-KV-KV-KV	KV-KV-KV-KVK-KVK
	VK-KV	VK-KV-KKVK	KV-KV-KV-KV	KV-KV-KV-KV-VK
	VK-KKV	VK-KV-KKV	KV-KV-KV-KVK	
	VK-KKVK	KV-KV-VK	KV-KV-KV-KV	
	KV-V	KV-KV-V	KV-KV-KVK-KVK	
	KV-VK	KV-KV-KVK	KV-KV-VK-KVK	
	KV-KV	KV-KVK-KVK	KV-KV-V-KV	
	KV-KVK	KV-V-KVK	KVK-KV-KV-KVK	
	KV-KKV	KV-KVK-KKVK	KV-KVK-KV-KV	
	KV-KKVK	KVK-KVK-VK	KV-KV-V-KVK	
	KVK-KV	KVK-KV-KV	KVK-KV-KV-KV	
	KVK-KVK	KV-KVK-KV	KVK-KV-KV-V	
	KVK-KKV	KVK-KV-KVK	KVK-KV-V-KVK	
	KVK-KKVK	KVK-KV-VK	KV-KVK-KV-KVK	
		KVK-KVK-KVK	KV-KVK-KV-VK	
		KVK-KVK-KKVK		
		KKV-KV-KVK		

2. Consonants do not stand alone as syllables,
3. Vowels can stand alone as syllables,
4. Consonant groups can only be at the beginning of the,
5. The second consonant in the consonant cluster generally consists of a few l, r, and y consonants,
6. *Ekasuku* phoneme arrangement is the basic pattern of two or more syllables.

2.3. Syllable Distortion

There are different ways to distort the original words (root words) and the affixes word morphemes.

a. The original word distortion

1. If in the middle of a word there are two vowels, the way of separating them should be between those two vowels.
   da-un <*daun*>
   leu-it <*leuit*>
2. If in the middle of a word there is a consonant between two vowels, the way to separation must be before the consonant.
   a-di <*adi*>
   sa-reng <*sareng*>
   wa-yang <*wayang*>
3. If in the middle of a word there is a consonant denoted by two letters, the letter renditions are not painted. the way to separation is before or after the letter cluster.
   di-nya <*dinya*>
   te-ngah <*tengah*>
4. If in the middle of a word there are two consonants that are not cluster (*rendon*) consonants, a way of separating between the two consonants.
   ang-klung <*angklung*>
   sar-ta <*sarta*>
   tek-tek <*tékték*>
5. If in the middle of a word there are two or more consonants is a *wianjana* mound, the consonant rendon is not painted. The disfigure before the wianjana mound.
   am-prok <*amprok*>

ga-plék <*gaplek*>
nam-bru <*nambru*>

b. The affixes word distortion

1. Prefixes, suffixes, and regular particles written tied to the original word, the way of separation must be independent.
   di-ba-wa <*dibawa*>
   da-har-eun <*dahareun*>
   da-han-na <*dahanna*>
2. Prefixes, suffixes, and the particles underwent a change of written form as follows.
   pa-na-nya <*pananya*>
   pa-nga-la <*pangala*>
   da-har-eun-a-na <*dahareunana*>
3. Infix of the written note with the word it comes from, distorting like distorting the original word.
   ba-ra-la <*barala*>
   da-la-har <*dalahar*>
   gu-meu-lis <*gumeulis*>
   pi-nang-gih <*pinanggih*>
4. Infix that changes its place in front of the word originally, considered equal to the next rank.
   ar-u-lin <*arulin*>
   al-u-dur <*aludur*>
   ra-cleng <*racleng*>
   um-a-mis <*umamis*>

2.4. Corpus Data

In this research, the corpus data was taken from [18] which contains Sundanese wave sound and a TSV file transcript. We use a male and female section in the TSV file then combine it to a single text file which produces 4213 sentences, 61383 tokens of words with duplication. For the input process, we use the words after the filter process. So, 3643 unique words will be used as testing data. Then, we manually validate for every word and we have 494 words with affixes form. This data will be used for algorithm implementation.

2.4. Design of Algorithm

The stemming design process in this study is by developing a technique for forming affixes based on [10] with a combination of affixes [14] as shown in Figure 2.

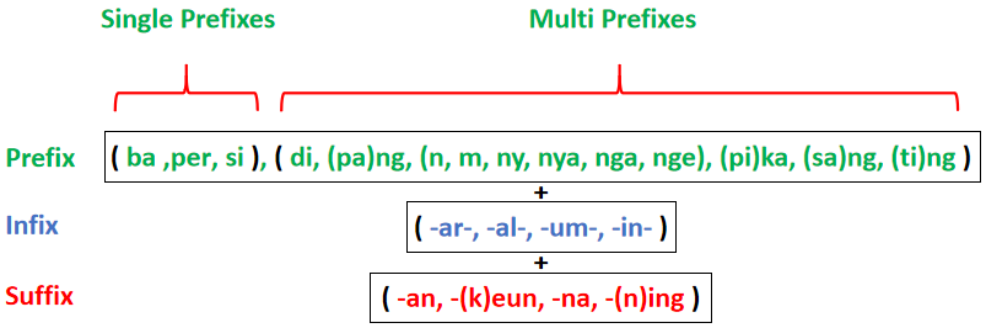

Figure 2. Affixes combination in Sundanese

Based on the rules of syllables and word derivatives in Sundanese, the design of the stemming algorithm with a syllable pattern can be seen in Figure 3.

p-ISSN: 2528-1682
e-ISSN: 2527-9165

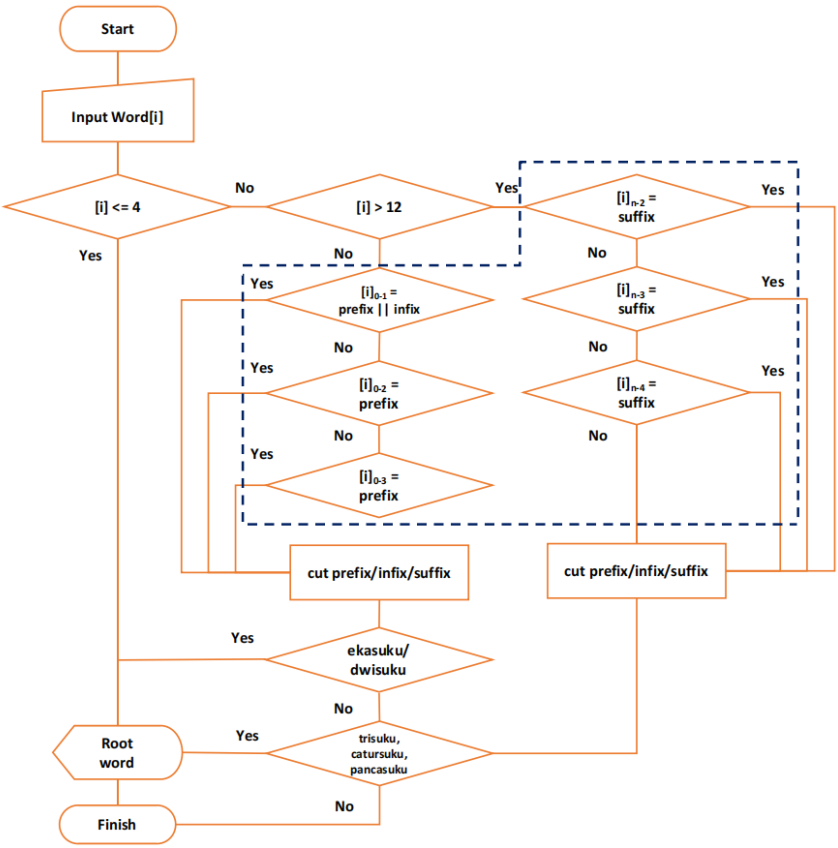

Figure 3. Stemming algorithm with syllable pattern

3. RESULTS AND DISCUSSION

After doing the research stages, in general, the use of algorithms with the syllable pattern feature can be applied to the stemming process, especially for the Sundanese language. We present the results and discussion for this study in the following points.

3.1. Result

In this study, the process of testing stemming algorithms using syllable patterns was performed on data sourced from [18] with 4213 sentences, 61383 tokens of words with duplication. Then filtered it with a result of 3643 unique words were obtained with a total of 494 affixes words. stemming results were obtained 440 syllable patterns (canonical) and 469 base words. Thus the stemming algorithm result using this syllable pattern has an accuracy of 89% with a base word result approach of 95%.

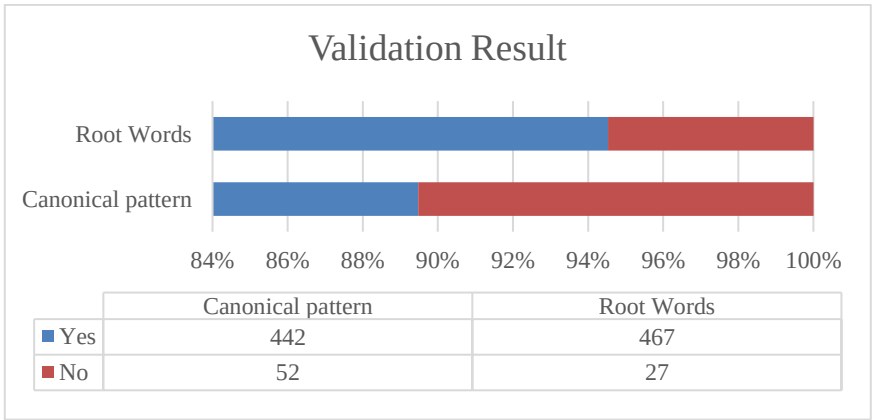

Figure 4. Stemming Process Result

3.2. Discussion

Syllable patterns have an important role in the word derivation in Sundanese. Sundanese syllables consist of *Ekasuku* (one of the syllables), *Dwisuku* (two of syllables), *Trisuku* (three of syllables), *Catursuku* (four of syllables), and *Pancasuku* (five of syllables). These patterns are conversions that represent vowels and consonants as well as become a root word or basic word in a word. a root word or basic word has a pattern that has at least one letter in the *Ekasuku* and a maximum of 12 letters in the five letters. Thus, the application of the stemming algorithm in this study is divided into three main processes following the syllable pattern. The canonical comparison process of a word with its syllables is illustrated in Figure 5.

Figure 5. The canonical form of the word

Stemming algorithms using syllable patterns can give good results in returning a word to a root word or root word. However, because Sundanese grammar is different from Indonesian, English, and others. then, several cases arise and become obstacles when the stemming process is run. First, the use of the Sundanese alphabet has differences, for example "/eu/ [ö]", "/ny/ [ñ]", and "/ng/ [ŋ]" [15] which can lead to confusion in the syllable conversion process. Second, the use of loan-word from Indonesian appears in the dataset [18], including for example "*Administratifna*", "*Efektifna*", "*Kesenian*", "*Perumahan*", and "*Perpustakaan*" which causes the stemming process unable to produce appropriate canonical patterns nor root words. Third, the use of nasal prefixes (allomorph) in Sundanese cannot be distinguished because the canonical form of syllables can have many possible words (non-deterministic) [12] so that an additional dictionary is needed as a reference for the proper use of nasal prefixes.

4. CONCLUSION

In this study, stemming using syllable patterns can be applied to return affixed words into the root or base word with an accuracy of 89%. The results of the execution of the test data resulted in 95% of the basic words. This simple algorithm has the advantage of being able to adjust the position of the syllable pattern with the word to be stemmed. Due to some data shortage constraints (typo, loan-word, non-deterministic syllable pattern), we can improve to increase the accuracy between 3-5%, by adjusting words and adding reference dictionaries. In addition, this algorithm has a drawback that causes the execution to be over-stemming.

ACKNOWLEDGEMENTS

The authors grateful for the research funded and supported by the Ministry of Education, Culture, Research, and Technology with parent contract number 065 /SP2H/LT/DRPM/2021 and derivative contract number 069/SP2H/RDPKR-MONO/LL4/2021. Institut Teknologi Garut with contract number 3/STTG/LPPM/SP-LPPM/II/2021.

5. REFERENCES

[1] E. Z. Arifin, “Bahasa Sunda Dialek Priangan,” Pujangga, vol. 2, no. 1, pp. 1–44, 2016.
[2] https://www.pikiran-rakyat.com/pendidikan/pr-01342765/bahasa-sunda-hadapi-tantangan-besar-pemerintah-lakukan-beragam-upaya. Accessed 1 July 2021.
[3] F. Amin and Purwatiningtyas, “Stemmer Bahasa Jawa Ngoko dengan Metode Affix Removal Stemmer (Rule Base Approach),” J. Teknol. Inf. Din., vol. 21, no. 1, pp. 16–24, 2016.
[4] N. Hidayatullah, A. P. Wibawa, and H. A. Rosyid, “Penerapan ECS Stemmer untuk Modifikasi Nazief & Adriani Berbahasa Jawa,” vol. 3, no. 3, pp. 343–348, 2019.
[5] R. Maulidi, “Modifikasi Metode Enhanced Confix Stripping,” Pros. Semin. Nas. FDI 2016, no. December, pp. 12–15, 2016.
[6] G. Ngurah, M. Nata, and P. P. Yudiastra, “Stemming teks sor-singgih Bahasa Bali,” Konf. Nas. Sist. Inform. 2017 STMIK, no. Agustus, pp. 608–612, 2017.
[7] M. Agus, P. Subali, C. Fatichah, and D. Informatika, “Kombinasi Metode Rule-Based Dan N-Gram Stemming Untuk A Combination Of Methods Rule-Based And N-Gram Stemming To Recognize Balinese Language Stemmer,” vol. 6, no. 2, 2019, doi: 10.25126/jtiik.201961105.

[8] D. Junaedi, O. Herlistiono, and D. Akbar, "Stemmer for 'Basa Sunda,'" Semin. Nas. ILMU Komput. Univ. DIPONEGORO, pp. 275–278, 2010.
[9] A. Purwoko, "Model Stemming Berbasis kamus untuk dokumen berbahasa sunda," INSTITUT PERTANIAN BOGOR, 2011.
[10] A. Ardiyanti Suryani, D. Hendratmo Widyantoro, A. Purwarianti, and Y. Sudaryat, "The rule-based sundanese stemmer," ACM Trans. Asian Low-Resource Lang. Inf. Process., vol. 17, no. 4, 2018, doi: 10.1145/3195634.
[11] A. Mirna, J. Asian, B. Nazief, S. M. M. Tahaghoghi, and H. Williams, "Stemming Indonesian : A confix-stripping approach," no. September 2018, 2007, doi: 10.1145/1316457.1316459.
[12] A. Purwarianti, "A non deterministic Indonesian stemmer," Proc. 2011 Int. Conf. Electr. Eng. Informatics, ICEEI 2011, no. October, 2011, doi: 10.1109/ICEEI.2011.6021829.
[13] P. Willett, "The Porter stemming algorithm: then and now," Program, vol. 40, no. 3, pp. 219–223, Jul. 2006, doi: 10.1108/00330330610681295.
[14] Y. Sudaryat, Tatabasa Sunda Kiwari. 2013.
[15] L. S. Faznur et al., "Komparasi fonem bahasa sunda dan bahasa indonesia dalam buku teks," Pena Literasi J. Pendidik. Bhs. dan Sastra Indones., vol. 2, no. 2, pp. 105–114, 2019.
[16] A. Djamaludin, M. Patoni, A. Sumantri, R. H. M. Koerdie, M. O. Koesman, and E. S. Adisastra, Kamus Sunda Indonesia. 1985., [Online]. Available: http://repositori.kemdikbud.go.id/2954/1/Kamus Sunda-Indonesia - %28449h%29a.pdf
[17] I. Baidillah et al., Direktori Aksara Sunda untuk Unicode, 1st ed. Dinas Pendidikan Provinsi Jawa Barat, 2008.
[18] K. Sodimana et al., "A Step-by-Step Process for Building TTS Voices Using Open Source Data and Framework for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese," in Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU), Aug. 2018, pp. 66–70, [Online]. Available: http://dx.doi.org/10.21437/SLTU.2018-14.