Improving Arabic Stemmer: ISRI Stemmer

Mochamad Gilang Syarief
Math Dept, UIN Sunan Gunung Djati,
Bandung, Indonesia
gilangmochamad39@gmail.com

Arief Fatchul Huda
Math Dept , UIN Sunan Gunung Djati,
Bandung, Indonesia
afhuda@uinsgd.ac.id

Opik Taupik Kurahman
Informatic Dept, UIN Sunan Gunung Djati,
Bandung, Indonesia
opik@uinsgd.ac.id

Wahyudin Darmalaksana
Ilmu Hadist Dept, UIN Sunan Gunung Djati,
Bandung, Indonesia
yudi_darma@uinsgd.ac.id

***Abstract*— Stemmer is used in several types of applications such as Text Mining, Information Retrieval (IR), and Natural Language Processing (NLP). Stemmer is a step used to process text data. The main task in the stemmer is to return the word-formation to the basic word (root or stem). ISRI Stemmer is one of the Arabic stemmers contained in the NLTK package. This study improves the weakness of the ISRI stemmer in processing words consisting of two letters. From the results of the experiment, these improvements increased the stemmer yield by 7.3%.**

Keywords—Stemmer, Text Mining, Information Retrieval, Natural Language Processing, ISRI stemmer

I. Introduction

Arabic is one of the Semitic languages, which also includes Hebrew, Aramaic, and Amharic. Arabic is a very important Semitic language because this language is the lingua franca from the Near East and North Africa [1].

It is estimated that there are around four hundred million Arabic speakers. Because this language is the language of religious instruction in Islam, many users from various countries at least have passive knowledge about the language [2]. Arabic is also one of the six official languages united nations (UN) and is based on 28 alphabet characters [3].

Arabic has a complex morphological structure. Morphology concentrates on the derivation of words and their effect on syntax. Most of the formations in Arabic come from the root word consisting of three letters [4] [5].

Stemmer is a method for getting root words from formations [6]. Stemmer can be defined as combining all variations of certain words into a single form called root or stem [7]. One process for finding the root word from the word-formation can be done by removing all affixes contained in the word [8]. In addition to the elimination of affixes, some words require changes in the shape of letters or arrangement of letters to get the root word.

The stemmer algorithm for Arabic is divided into two groups. First, a light stemmer, which is the process of removing affixes (prefixes, infixes, and suffixes) to get the root word [4]. Examples of light stemmer are Snowball stemmer [9] and ARLStem [10]. Second, a heavy stemmer, which is the process of eliminating affixes, and changing some letters in words to get the root word [4]. Examples of heavy stemmers are Khoja stemmer [11], ISRI stemmer [12], Ghwanmeh stemmer [4] and Al-Kabi stemmer [6]. Another term for heavy stemmer is root-extraction stemmer [12].

Some changes in letters in a heavy stemmer can be done in two ways. First, with word pattern matching. The pattern in question is a morphological pattern in Arabic. Stemmers who use this method are ISRI stemmer and Al-Kabi stemmer. Second, matching the root word dictionary. The dictionary in question is a collection of root words that are used as a comparison of the stemmer results. Stemmers who use this method are Khoja stemmer.

Snowball stemmer, ARLStem, and ISRI stemmer are in the NLTK package. NLTK is a platform used to create Python programs to process data in the form of language [13]. Python is a simple but powerful programming language with excellent functionality for processing linguistic data [14].

Of the three stemmers above, ISRI stemmer returns form words to become root words using patterns. From the experiments that the author did, this stemmer returns the same word for words consisting of two or three letters.

For stemming three-letter words, the ISRI stemmer returns the original word (not processing). This has two possibilities. First, the returned word is correct as a root word. Second, the word is wrong because it's not the root word. For example, the word (رَبُّكَ), the word return (ربك) that should be returned is (ربب).

In this study, proposed improvements from the ISRI stemmer. The improvements made focus on handling word conditions consisting of 2 letters. Additional conditions are carried out after the ISRI stemmer processes the word.

II. Related Work

The complex morphological structure of the Arabic language has resulted in some researchers being interested in improving the processing of Arabic words to obtain efficient stemmer results.

Several studies have been carried out to develop Arabic language stemmers such as Khoja stemmer made by Shereen Khoja et al (1999) for Arabic text documents, by implementing the algorithm into the java programming language into a desktop-based application and using the word root dictionary to extract words into forms root word [11].

Kazem Taghva et al (2005) studied the making of stemmer (root-based) using patterns to extract words into basic words. They apply root extraction stemmer for Arabic which is similar to Khoja stemmer but without a word root dictionary [12]. The algorithm in this research is used in the ISRI stemmer.

Sameh Ghwanmeh et al (2009) presents Arabic root-based algorithms based on morphological patterns. This algorithm has been tested using thousands of Arabic words taken from

978-1-7281-4796-3/19/$31.00 ©2019 IEEE

the corpus 242 abstracts from the Proceedings of the Saudi Arabian National Computer Conference. The results obtained show that the algorithm extracts the correct root with an accuracy of up to 95% [4].

Mohammed N. Al-Kabi et al (2015) applied light stemmer and heavy stemmer to Arabic words to extract the roots of triliteral words. Data sets consisting of 6081 Arabic words originating from the Arabic literal verb were built to evaluate the Al-Kabi stemmer algorithm against Khoja stemmer and Ghwanmeh stemmer. The data used include single, multiple, and plural Arabic words (nouns and verbs) originating from the roots of the triliteral Arabic words. The evaluation results for Al-Kabi stemmers are superior in some cases compared to the other two stemmers [15].

Kheireddine Abainia et al (2016) proposed the design and implementation of a light stemmer based on several new rules for removing prefixes, suffixes, and infixes without using dictionary words or patterns. The researcher's main objective is to reduce the data dimension and rearrange words that are relatively morphological and semantic [10].

Aditya Hanif Utama et al (2018) in his research developed a computer-based system for stemming development in the Qur'an using the Shereen Khoja stemmer algorithm. The development of stemming in the Qur'an is an important jo because it supports the Sharaf classification to understand the meaning of the word [6].

Some researchers present examples of Arabic stemmer algorithms and their effectiveness. Some researchers claim the high accuracy of each proposed algorithm. However, the lack of source codes and data sets resulted in testing is unable to be done.

III. Methodology

The ISRI stemmer is built on algorithms: Arabic stemmer without a root dictionary developed by the Information Science Research Institute (ISRI) of the University of Nevada Las Vegas, the USA using an algorithm made by Kazem Taghva, Rania Elkhoury, and Jeffrey Coombs (2005). The algorithm from ISRI Stemmer is as follows:

1. Removes diacritics that represent vowels,
2. Normalization of Hamza (Changing letters ؤ ,ء, and ئ, into letters أ),
3. Delete two-letter and three-letter prefixes in the word sequence,
4. Delete connecting letters و if there is a letter و in the prefix of the word,
5. Alif normalization (Changing letters أ ,ٱ, and إ, to letters ا),
6. Stemmer returns the same letter if the word 3 letters. And return the same word if the word is ambiguous,
7. Consider 4 cases depending on the length of the word:

- 4 letter length,
- 5 letter length,
- 6 letter length,
- 7 letter length.

The ISRI stemmer written by Hosam Algasier uses an algorithm: Kazem Taghva, Rania Elkhoury, and Jeffrey Coombs (2005). There are additional adjustments made by Hosam Algasier to improve the algorithm:

1. Add 60 stop words,
2. Add a pattern (تفاعيل) to the ISRI stemmer pattern data set,
3. Step 2 in the original algorithm is to normalize all Hamza. This step is discarded because it increases word ambiguity and changes the original word root.

The improvements made in this study were by adding conditions after the word was processed by the ISRI stemmer. There are several rules added to overcome words that consist of two letters. First, the word by word will be checked in a collection of double words. Then the word will be checked whether it includes words with last weak letters, first weak letters, or middle weak letters. The flowchart of the program built is shown in Fig. 1.

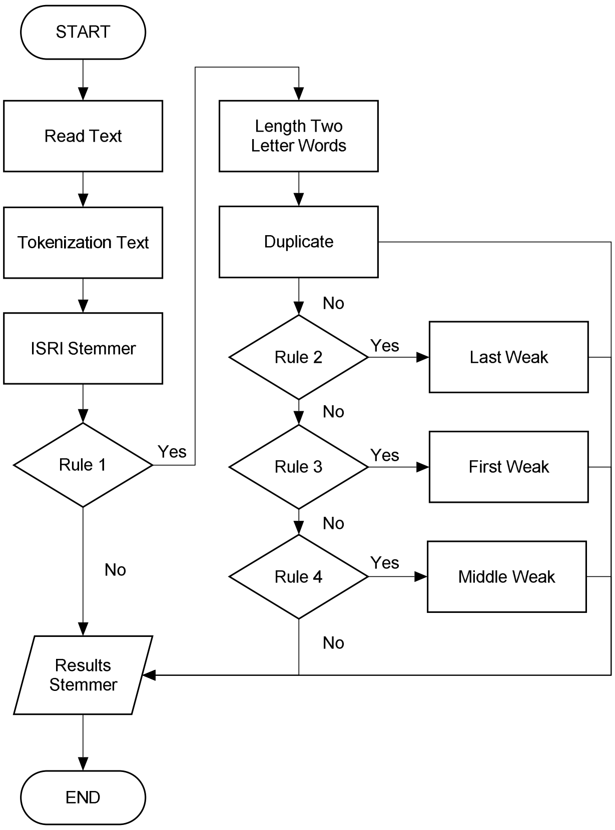

Fig. 1 Revise of ISRI Stemmer

IV. Results and Evaluation

A. Data

The data sets used in this study are sourced from the Quranic Arabic Corpus [16], an annotated linguistic source that shows Arabic grammar, syntax, and morphology for each word in the Qur'an the chapter Al-Fil until An-Nas and 35 two-letter words that have been collected.

There are several variables used in Table I. "Name" represents the name of the chapter from the Qur'an. "WTR" represents the number of words that have a word root in the letter, "WTNR" represents the number of words that have no word root in the letter, and WT is the total word in the letter.

TABLE I. DATA SET OF CHAPTER AL-FIL TO AN-NAS

No	Name	Word			Number of verses
		WTR	WTNR	WT	
1	An-Nas	16	4	20	6
2	Al-Falaq	15	8	23	5
3	Al-Ikhlas	10	5	15	4
4	Al-Masad	17	6	23	5
5	An-Nashr	16	1	17	3
6	Al-Kafirun	12	14	26	6
7	Al-Kautsar	7	3	10	3
8	Al-Maun	14	11	25	7
9	Al-Quraisy	12	5	17	4
10	Al-Fil	18	5	23	5
Total		137	62	199	48

B. Results and Evaluation

Additional conditions are used to deal with the condition of two-letter words. In conditions that are added, there is a collection of root words for two-letter words, which are stored in text format. This file is a collection of words that have been collected from Shereen Khoja's research [11]. However, there are some adjustments to the data used. Adjustments made are additions (جل) ,(تم) ,(فم) ,(فر) ,(حب) ,(مر) ,(مد) ,(دل) ,(جب) (حج), (شر) and deletion of words (قل) in a collection of root words.

The following Table II, III and IV shows the results of the ISRI stemmer process before and after improvements in duplicate and fi'il mu'tal words.

TABLE II. RESULTS OF ISRI STEMMER ISIM DUPLICATE

Word	ISRI	Rev_ISRI	Root	Categories
رَبِّ	رب	ربب	ربب	Isim
شَرِّ	شر	شرر	شرر	Isim

TABLE III. RESULTS OF ISRI STEMMER FI’IL MADHI DUPLICATE

Word	ISRI	Rev_ISRI	Root	Categories
حَجَّ	حج	حجج	حجج	Fi’il Madhi
سَرَّ	سر	سرر	سرر	Fi’il Madhi
مَرَّ	مر	مرر	مرر	Fi’il Madhi
فَرَّ	فر	فرر	فرر	Fi’il Madhi
مَدَّ	مد	مدد	مدد	Fi’il Madhi
دَلَّ	دل	دلل	دلل	Fi’il Madhi
جَبَّ	جب	جبب	جبب	Fi’il Madhi
حَبَّ	حب	حبب	حبب	Fi’il Madhi
تَمَّ	تم	تمم	تمم	Fi’il Madhi
جَلَّ	جل	جلل	جلل	Fi’il Madhi

From the stemmer results carried out on 35 words consisting of two letters, only 30 words can be returned by the stemmer into the correct word root consisting of three letters. While the other 5 words cannot be returned to the correct three-letter root.

TABLE IV. ISRI RESULTS STEMMER FI'IL MU’TAL

Word	ISRI	Rev_ISRI	Root	Categories
قُلْ	قل	قول	قول	Fi’il Amar
هَبْ	هب	وهب	وهب	Fi’il Amar
قُمْ	قم	قوم	قوم	Fi’il Amar
خُنْ	خن	خين	خين	Fi’il Amar
تُبْ	تب	توب	توب	Fi’il Amar
صُنْ	صن	صون	صون	Fi’il Amar
طُفْ	طف	طوف	طوف	Fi’il Amar
دُرْ	در	دور	دور	Fi’il Amar
زُرْ	زر	زير	زير	Fi’il Amar
صُمْ	صم	صوم	صوم	Fi’il Amar
عُدْ	عد	عود	عود	Fi’il Amar
ضَعْ	ضع	وضع	وضع	Fi’il Amar
دَعْ	دع	ودع	ودع	Fi’il Amar
رَعْ	رع	ورع	ورع	Fi’il Amar
ضَحْ	ضح	وضح	وضح	Fi’il Amar
قَعْ	قع	وقع	وقع	Fi’il Amar
نَعْ	نع	ينع	ينع	Fi’il Amar
نِلْ	نل	نول	نول	Fi’il Amar

To evaluate the results of ISRI stemmer improvements, the following are presented in Table V, which are the stemmer results of several short chapters in the Qur'an.

TABLE V. ISRI STEMMER RESULTS IN SHORT LETTERS

Name	ISRI			%	Rev_ISRI			%
	T	F	WT		T	F	WTR	
An-Nas	4	12	16	25.00%	6	10	16	37.50%
Al-Falaq	8	7	15	53.33%	13	2	15	86.67%
Al-Ikhlas	5	5	10	50.00%	6	4	10	60.00%
Al-Masad	8	9	17	47.06%	8	9	17	47.06%
An-Nashr	7	9	16	43.75%	7	9	16	43.75%
Al-Kafirun	11	1	12	91.67%	12	0	12	100.00%
Al-Kautsar	2	5	7	28.57%	2	5	7	28.57%
Al-Maun	5	9	14	35.71%	5	9	14	35.71%
Al-Quraisy	7	5	12	58.33%	8	4	12	66.67%
Al-Fil	14	4	18	77.78%	14	4	18	77.78%
Total	71	66	137	51.82%	81	56	137	59.12%

There are several variables used in Table V. Like the "Name" in question, namely the name of the chapters from the Qur'an. "T" represents the number of words that are correctly extracted into the root word. "F" represents the number of words that are incorrectly extracted into the root word, and "WTR" represents the number of words that have a root word in the data set. The percentage is calculated by dividing the variable T with WTR.

From the results of the experiment, the improvements made increased the results of the correct word return on some of the letters tested. The percentage of total stemmer increased by 7.3% from the percentage before revises which were 51.82% to 59.12%. Fig. 2 shows the results of the number of words correctly extracted into the root word, before and after the ISRI stemmer revise.

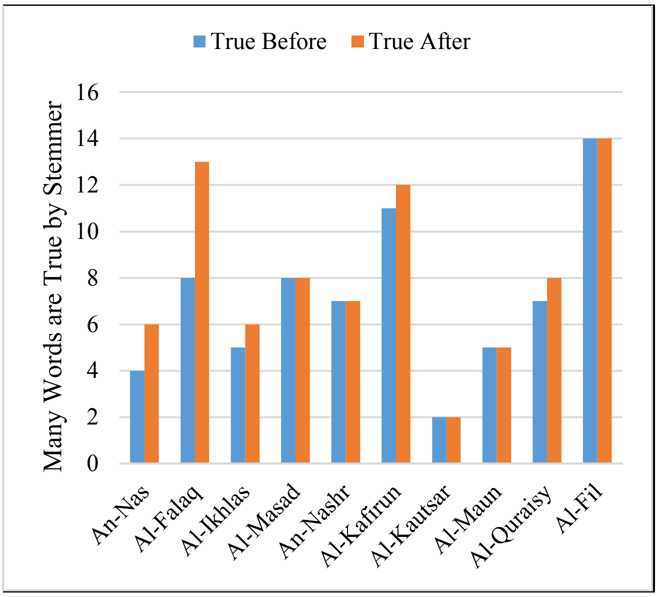

Fig. 2 ISRI Results of Stemmer Before and After Revise

In other cases, alif normalization is useful for the data processing carried out by ISRI stemmers. Normalization is meant to change the letter (ٱ) to letter (ا). This is useful because in the ISRI stemmer process when the word has a letter (ٱ) at the beginning of the word, the word cannot be processed. Whereas in the data set used there is a letter (ٱ) which is part of the word.

Changing the letter (ٱ) to letter (ا) in the word (ٱلرَّحْمَٰن) is proven to simplify the stemmer process carried out by the ISRI stemmer. Because changing letters produces stemming processes can be done on the word (ٱلرَّحْمَٰن) where the word comes from the basic word (رحم).

V. Conclusion and Future Work

From the results of the experiment, improvements made increased stemmer yield. The results of the improvement are proven to be able to overcome the problem for the stemmer process in several words consisting of two letters. For example, the word (رَبِّ) ,(شَرِّ) ,(قُلْ) which has been returned to become the root word (ربب), (شرر) and (قول). The improvements made increased by 7.3% on the total correct words returned by the stemmer. In the future, a larger set of data is needed to observe the results of improvements in the handling of words consisting of two letters. In other cases, the normalization of alif is useful to simplify the process carried out by the stemmer.

Acknowledgment

We would like to say thank you for the Research and Publication Centre of UIN Sunan Gunung Djati Bandung that gives the full support and fund for this publication research.

References

[1] A. D. Rubin, “A Brief Introduction to the Semitic Languages,” Gorgias Handbooks; 19, Gorgias Press, 2010.

[2] M. Mustafa, A. S. Eldeen, S. Bani-Ahmad, A. O. Elfaki, “A Comparative Survey on Arabic Stemming: Approaches an Challenges,” *Intelligent information Management, 2017, 9,* pp. 39-67, 2017.

[3] M. N. Al-Kabi, Q. A. Al-Radaideh, K. W. Akkawi, “Benchmarking and assessing the performance of Arabic stemmer,” *Journal of Information Science,* 37(2), pp. 111-119, 2011.

[4] S. Ghwanmeh, R. Al-Shalabi, S. Rabab’ah, G. Kanaan, “Enhanced Algorithm for Extracting the Root of Arabic Words,” *International Conference on Computer Graphics, Imaging and Visualization*, 2009.

[5] A. Al-Omari, B. Abuata, “Arabic Light Stemmer (ARS),” *Journal of Engineering Science and Technology*, 9(6), pp.702-717, 2014.

[6] M. N. Al-Kabi, S. A. Kazakzeh, B. M. A. Ata, S. A. Al-Rababah, and I. M. Alsmadi, "A novel root based Arabic stemmer," Journal of King Saud University – Computer and Information Sciences, vol.27, no.2, pp. 94-103, 2015.

[7] T. M. T. Sembok, B. M. A. Ata, Z. A. Bakar, “A Rule-Based Arabic Stemming Algorithm,” *Proceedings of the European Computing Conference*, 2011.

[8] H. Widayanto, A. F. Huda, “Comparison Nazief Adriani and CS Stemmer Algorithm For Stemm Real Data” *e-Proceeding of Engineering*: Vol.4, No.3, pp. 5215-5222, 2017.

[9] A. Chelli, “Assem’s Arabic Light Stemmer (BETA)”. [Online]. Available: https://www.arabicstemmer.com/ [Accessed: May 10, 2019].

[10] K. Abainia, S. Ouamour, H. Sayoud, “A novel robust Arabic light stemmer,” *Journal of Experimental & Theoretical Artificial Intelligence*, 2016.

[11] S. Khoja, R. Garside, “Stemming Arabic text.” Computing Department, Lancaster University, Lancaster, UK, 1999.

[12] K. Taghva, R. Elkhoury, J. Coombs, "Arabic Stemming Without A Root Dictionary," In: *International Conference on Information Technology: Coding and Computation (ITCC 2005)*, pp. 152-157, 2005.

[13] S. Bird, L. Tan, “Natural Language Toolkit .” [Online]. Available: https://www.nltk.org/ [Accessed: May 10, 2019].

[14] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python, Vol. 43, 2009.

[15] A. H. Utama, M. A. Bijaksana, A. F. Huda, “Pengembangan Sistem Berbasis Komputer Untuk Pembangunan Stemming pada Al-Quran Menggunakan Algoritma Shereen Khoja Stemmer,” *e-Proceeding of Engineering*: Vol.5, No.2, pp. 3657-3669, 2018.

[16] Leeds University, “Quranic Arabic Corpus”. [Online]. Available: http://corpus.quran.com/ [Accessed: May 10, 2019].