IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 3, Ver. VI (May – Jun. 2015), PP 25-31
www.iosrjournals.org

A Word Stemming Algorithm for Hausa Language

Muazzam Bashir[1], Azilawati Binti Rozaimee[2], Wan Malini Binti Wan Isa[3]
[1, 2, 3] *(Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin Tembila Campus 22200 Besut Terengganu, Malaysia)*

Abstract: *Hausa, a highly inflected language, needs a worthy stemming approach for efficient information retrieval (IR). However, there is a limited or unavailable study to stemming in the language. Stemming refers to the systematic way of reducing a word to its base or root form. It is a crucial aspect in the field of natural language processing (NLP) such as text summarization and machine translation. As such, this study inspirationally presents an automatic word stemming system for Hausa language with a view to contributing to the field of electronic text processing, as well as NLP, in general. The proposed method is a modification of Porter's algorithm to fit Hausa morphological rules. The system has an accuracy of 73.8% for implementation with 2573 words extracted from four different articles from Hausa Leadership newspaper. If immensely improved over time (employing more exceptional cases in future work), it would inspire the development of more tools for the language. Hence, the language would rapidly adopt the advancement in technology.*

Keywords: *Hausa language, Information retrieval, Natural language processing, Stemming*.

I. Introduction

A stemming algorithm refers to a computational method of reducing all words having the same root-form to the same form by stripping affixes or infixes attached to those words. Stemming programs are also called stemming algorithms or stemmers [1]. One of the advantages of stemming is improving the performance of information retrieval (IR) by providing morphological variants of key terms with a view to matching the documents with the given query. The words in the request might have a lot of variants [2]. For instance, a user might have "stemming" in their query and would likely be interested in "stemmed" or "stem". Conflation is the process of matching morphological variant terms. Conflation is either manual, employing some regular expressions or automatic – using stemmers. Moreover, automatic conflation can be language dependent or language independent, each of which has its advantages over the other. Several studies are still going on both of the two conflations. Several studies that include English stemmer in [1], UniSZA stemmer for Malay [3], etc., are language dependent studies. The improvement in technology drew the attention of many researchers to conduct studies for language neutral stemming [4] [5].

Stemming is inappropriate in two ways namely: over-stemming and under-stemming [2]. Over-stemming a word refers to too much removal of its part by stemming process, and this has a serious effect on IR as it would lead to the retrieval of non-relevant document due to conflation of unrelated terms. Under-stemming a word means removing too little part of a word, and this can prevent the retrieval of relevant documents.

This study focuses on language dependent conflation. It is the process of using a given language set of standard rules of inflection or derivation of a similar group of words to reduce any given word, which belongs to such groups, to its stem. Fortunately, the need for a corpus of the target language is very limited as compared to Language Independent Approach. Furthermore, stemming is also a crucial aspect in the field of automatic text summarization especially the abstractive form – that requires a sentence and word modification. It is believed that enhancing the rules for a particular language is significantly reducing errors in an automated conflation [3].

There are very limited or unavailable studies that attempted to develop a stemming algorithm for Hausa language despite the fact that Hausa language is an International language widely spoken in West Africa with estimated speakers of over 52 million. It is a very popular language in many African countries such as Nigeria, Niger, Togo, Ghana and Mali. Nigeria has at least 40 million Hausa speakers. It happens to be the top language in Africa [6]. It has two forms of writing system namely: Ajami and the Boko system. The Ajami uses modified Arabic alphabets and became acknowledged in early 17th century. While the Boko system, Latin-based, was introduced in 1930s by the colonial administration. The Boko system is the most prevalent in the growing printed Hausa literature such as books, novels, newspaper and plays. The language also has different dialects such as "Kananci", "Katsinanci" and "Zazzaganci". The "Kananci" (from Kano) is the standard in Hausa literature and it is adopted by the international Hausa news broadcasters such as BBC Hausa and VOA Hausa. Several high degrees (Masters and Ph.D.) in Hausa language are offered overseas universities in the US, UK, and Germany [7].

In this study, a word stemming system based on Hausa language morphological rules was proposed which modified the Porter's stemming algorithm [12] and also used some rules proposed in [3] accordingly.

II. Literature review

There are four categories of Automated Conflation based on their approaches namely: Successor Variety, Affix Removal, n-gram and Table Lookup methods. Fig. 1 below shows the two forms of conflation as mentioned earlier and stemmers' approaches [2].

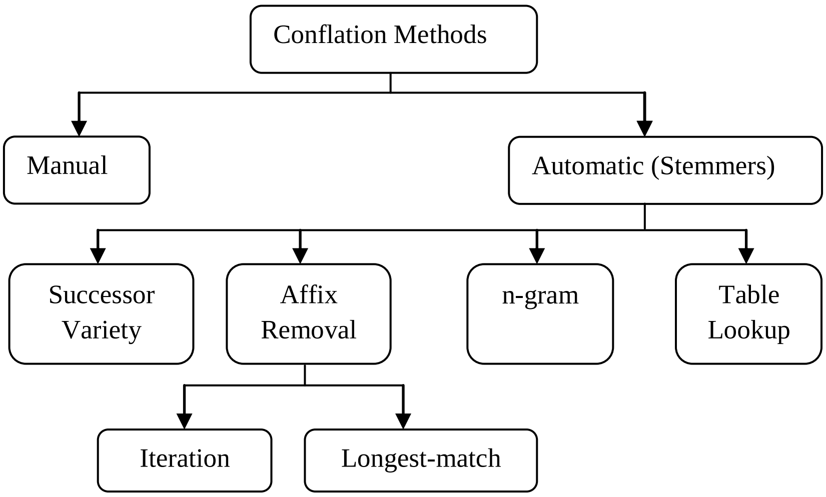

Figure 1: Conflation methods

1.1 Successor Variety

Successor variety employed frequencies of letter series, [8], to segment a word into its stem by utilizing a word corpus in the process of stemming. For instance, consider a word corpus: "kanwa", "Katanga", "kunya", "kubewa" and "kango". To find the successor variety of the word "kaya" then the initial letter of the target word "k" matches five words with distinct two second letters. The next prefix "ka" matches three words with two different third letters. The next "kay" matches no words, likewise the complete word "kaya". Hence, the word "kaya" has the successor variety of zero. Having determining the Successor Varieties for a test word taking peaks into consideration, then the detection and segmentation of the word preceded.

2.2 Affix Removal

In this approach, the suffix, prefix and infix are stripped from the words with a view to producing their stems. This method has been widely used since the last five decades and still common in nowadays stemmers. The method is mainly on two principles namely: Iteration and longest-match [1].

2.2.1 Iteration

This approach uses affix stripping with looping around the given order-classes and without repeating a match of the affix within each particular order-class.

2.2.2 Longest-match

Longest-match approach maintains that within a given set of endings, should more than one ending produce a match then the removal of the longest one followed.

2.3 N-Gram

A method of conflating terms called the shared digram proposed by [9]. A digram refers to a couple of successive letters. A research in [10] generalized the digram method and called it n-gram method. That is after testing for trigram, tetra-gram, etc.

In this approach, association measures are determined by the couple of terms based on shared distinct digrams. For instance, the term "dangantaka" (relationships) and "dangana" (to rely on) can be broken into digram as follows:

dangantaka =>	da	an	ng	ga	an	nt	ta	ak	ka
unique digram =	ak	an	da	ga	ka	ng	nt	ta	
dangana =>	da	an	ng	ga	an	na			
unique digram =	an	da	ga	na	ng				

The term “dangantaka” has nine digrams eight of which are unique while “dangana” has six digrams five of which are unique. And the two words share four digrams. After the determination of the unique digrams, then the similarity measure can be evaluated using Dice’s coefficient [2] as follows:

$$S = \frac{2C}{P + Q}$$

P is the number of unique digrams of the first term. Q is the number of unique digrams of the second term while C is the number of unique digrams shared by P and Q.

Hence, for the above example, we have: P=8; Q=5; and C=4 and therefore S = 2*4/(8+5) = 0.62. Similarly, this is calculated for all pairs of words in the database and then clustered into groups. This similarity measure gives a hint where the stem lies in every pair.

2.4 Table Lookup

Table Lookup is the process of storing all the words with their corresponding stems in a stand as reported in [11]. All the terms in both the query and indexes are stemmed through lookup from the table. The major obstacles to this approach include extensive use of a particular language to the stems of each word; it is also a time-consuming process and requires a significant storage overhead for such tables.

2.5 Related Work

The earliest published studies in the field of stemming emerged in the 1960s and all of which are English language dependent. The first published research in stemming algorithm, [1], is for English language. It was a context sensitive based on longest-match rule developed for use in a library to maximize the performance of IR. Later, the second published, Porter stemmer, proposed in [12] that concentrated on suffix removal in English words. The system was reasonably straightforward, fast, having small lines of codes and also better than the existing methods. The studies in [1] and [12] served as the basis for the later studies particularly in other languages. Some of the studies that aimed at developing a stemming algorithm in other languages include Wolaytta text stemmer proposed in [13] which was a prototype context sensitive iterative stemmer. The system has an accuracy of 90.6% on training using 3537 Words. Moreover, the system has an accuracy of 86.9% with unseen 884 words. A light weight stemmer, a rule-based, for Urdu text [14] ensures very abundant exceptional cases before undergoing stemming process. The system is quite enough to be effective in IR.

A UniSZA Stemmer is an enhanced system developed in [3] for Malay language that was determined to improve the processing speed of the previous methods. The study proposed seven simple rules that would disregard the dependency on Malay dictionary even though it also relies on Malay root words. Subsequently, the system has superior performance over the existing systems. Fortunately, this study used the idea of suffix stripping in Porter stemmer [12] as a hint due to the presence of infixes in Hausa words. In addition, the study also considered some of the rules proposed in UniSZA stemmer [3] all with a view to fitting the Hausa language stemming approach.

III. Proposed method

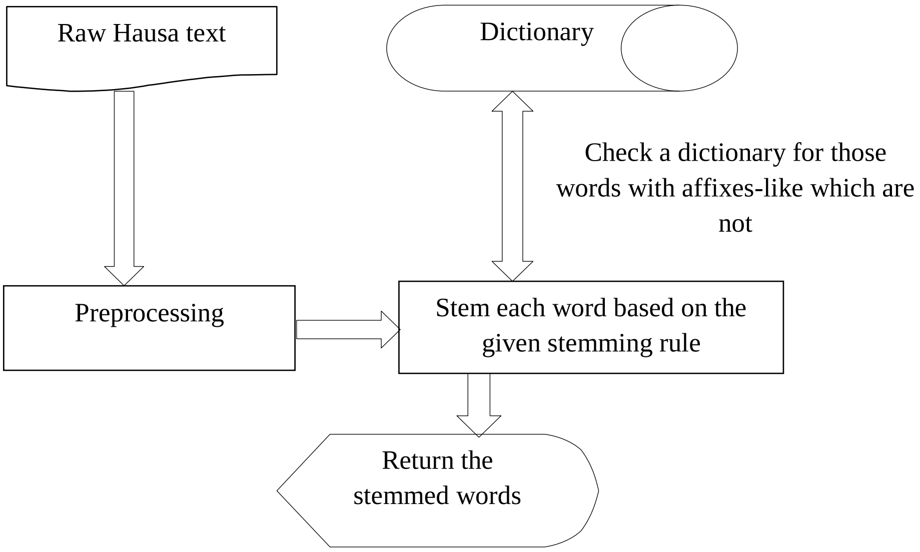
Raw Hausa text
Dictionary
Check a dictionary for those words with affixes-like which are not
Preprocessing
Stem each word based on the given stemming rule
Return the stemmed words

Figure 2: Hausa stemmer framework

This study proposed a system for Hausa stemming which is a context sensitive and iterative in nature. The study modified the algorithm in [12] and used some rules in [3], to handle exceptional cases that occur in Hausa language. The study thoroughly analyzed the Hausa word morphology presented in [15], [16], [6] and [7]. For this reason, this study developed the Hausa stemmer that utilizes the following rules check length, Stopword list, masculine & feminine gender marker, prefix list, suffix list, infix rule and check a dictionary. The proposed system was designed using Java programming language. It starts by reading the Hausa text from the collected article samples saved in notepad file. Subsequently, the file undergoes preprocessing, which involves the removal of punctuations, removal of words starting with a digit and changing the entire text case to lower case as well as tokenizing the text. For every word in the text, the stemming rule is applied and checks a dictionary where applicable. Finally, the system returns the stemmed words. The fig. 2 above demonstrates architecture of the system.

3.1 Check a Dictionary

Hausa words, sometimes, have complexity in the Affix stripping process. There exist many words that have prefix-like, suffix-like and sometimes both of which are neither suffix nor prefix. For instance, the word "babur" (motorbike) has prefix-like of "ba" and suffix-like of "r". In this case, a dictionary can be used to tackle the problem.

3.2 Prefix list

This list contains the morphemes that are used with stems to derive new words in Hausa language. Each of them would be checked on any given word for stemming by the system. Some of them are "ba", "ma", "mai", "yan", etc.

3.3 Suffix list

This list contains the morphemes that are attached to stems for formation of new words in Hausa language. For example, the word dakakku (pounded) was derived from daka (pound); in this case "kku" is a suffix. Other suffixes include wa, iya, uwa, anya, etc. But in some instances, Hausa new words are formed by changing the final vowel of a word. For example, derivation of kujer**u** (chairs) from kujer**a** (chair), the character 'a' changed to 'u'. A Hausa word sometimes contains gender marker and a pronoun attached to it. Therefore, that attachment is regarded as a suffix in this system. Consider the word motarsa (his car), the feminine gender marker 'r' and the pronoun "sa" (his) are attached to the "mota" (car). Other suffixes include nka, nshi, rta, etc.

3.4 Infix rule

Hausa morphology has complex infixes rules. These rules stripped words having unlisted suffixes in the suffix list with a view to decreasing the dimensions of the suffix list. For instance, the plural word lambobi (numbers) derived from lamba (number) belongs to the **o…i** rule: $C_1V_1 \ldots C_{n-1}V_{n-1}C_nV_n$ having where $C_{n-1} = C_n$ and $V_{n-1}=\mathbf{o}$, $V_n=\mathbf{i}$**.** If this rule is correct, then $V_{n-1}C_nV_n$ would be replaced by a character '**a**'. More Hausa words belonging to this instruction include harkoki (activities), hanyoyi (ways), kofofi (doors), bindigogi (guns), etc.

Another rule is **e…a…i** rule: For instance, the word kar**esa**n**i** (heifers) derived from karsana (heifer). In this rule, the '**e**' is dropped, and the '**i**' is changed to '**a**' to get the stem back. The words that belong to this rule include karefasi, tarewadi, etc.

3.5 Check a length

In this function, the lengths of the words are determined, and only those words exceed the given threshold value would undergo the stemming process. This study sets the limit to be three.

3.6 Stopword list

This list stored all the Hausa words that frequently occur in a text. The system discards stopwords upon encountered. The system moves to the next word if it meets a word that belongs to the stopword list. These words include "a", "wannan", "nan", "can", "cewa", etc.

3.7 Masculine & Feminine gender marker

In Hausa texts, the definite and indefinite articles are attached to the words using 'r' for feminine and 'n' for masculine. Feminine marker is attached to feminine words while masculine to masculine words. Therefore, this gender marker has to be stripped to stem the particular word. And there are some words that are exceptional, e.g. "babur".

Here is the proposed algorithm:

```
1.  Read the raw text
2.  Tokenize the text into words
3.  READ the word to be stemmed.
4.  IF the word contains letters only
            Go to 5
            ELSE
                    Go to 10
5.  IF the size of the word is greater than three
            IF the word is in the Stopword List
                    Go to 10
            ELSE
                    Go to 6
    ELSE
            Go to 7
6.  IF the word ends in ‘r’ or ‘n’
            IF the word ends in ‘r’
                    IF ‘u’ or ‘i’ comes before the ‘r’
                            Go to 7
                    ELSE
                            Remove the ‘r’ and Go to 8
            ELSE
                    Remove the ‘n’ and Go to 8
7.  RETURN the word and RECORD it in a stem dictionary and Go to 10
8.  IF the word starts with any of the prefixes in the Prefixes list
            Remove the prefix and
            CHECK Dictionary IF the resulting word exists
                    Go to 7
            ELSE
                    IF the word ends with any of the Suffixes list
                            Remove the suffix and Go to 7
                    ELSE
                            Restore the Prefix and Go to 9
    ELSE
            IF there exists immediate duplicate of first three letters in the word
                    Remove the first three letters and Go to 7
            ELSE
                    IF the word is a double word
                            Return the first word & Go to 7
                    ELSE Go to 9
9.  IF the word ends with any of the Suffixes in the Suffixes list
            Remove the suffix and
            CHECK Dictionary IF the word exist
                    Go to 7
            ELSE
                    IF the word has any infix rule
                            Remove the infix and Go to 7
                    ELSE
                            Go to 7
    ELSE
            IF the word has any infix rule
                    Remove the infix and Go to 7
            ELSE
                    Go to 7
10. IF the end of the text is not met Go to 3
            ELSE Go to 11
11. Stop
```

IV. Result and discussion

As mentioned earlier, with a very limited research in stemming for Hausa, then the proposed system's outcome was judged by a linguistic expert in the language. The study used a collection of 2573 words from four different articles in Hausa Leadership newspaper to evaluate the system.

Out of 2573 words, 966 words are detected as Stop words. The remaining words are 1607 and only 741 discrete words. Hence, 741 discrete words were used in the system on which 547 words were stemmed correctly. On the other hand, 110 and 84 words were over-stemmed and under-stemmed respectively as stemming error. The system has an accuracy of 73.8%. Table 1 provides a summary of this evaluation.

Table 1: Highlights of the System Outcome

Words	Expected Output	System Output (Stemmed Words)	Remarks
asibitin (the hospital)	asibiti	asibiti	Stemmed properly
musulmai (Muslims)	musulmi	musulmi	Stemmed properly
sa'o'i (hours)	sa'a	sa'a	Stemmed properly
sojojin (the soldiers)	soja	soja	Stemmed properly
shugabannin (the leaders)	shugaba	shugaba	Stemmed properly
wayoyin (the phones)	waya	waya	Stemmed properly
kare (dog)	kare	ka	Over-stemmed
kazanta (mess)	kazanta	kaza	Over-stemmed
kunne (ear)	kunne	ku	Over-stemmed
mummunan (the ugly)	muni	muna	Under-stemmed
sabuwar (the new)	sabo	sabu	Under-stemmed

The Table 1 above indicates that even with the use of the Hausa Dictionary, the system still encountered stemming error. When the system gets the word "kare" it is expected to return "kare" as the stem but the system applied the stemming rule and checked in the dictionary and found the output "ka" is valid. Same goes with the words kazanta with a stem "kaza" and "kunne" with "ku". One way of solving this problem, though quite expensive, is to find all the words with this ambiguity and put them into an exceptional list.

The problem of under-stemming has to do with the complexity of the Hausa morphology as it has so many inconsistent rules. Hausa verbs ending in 'a' are the root forms, and all others are derived form. The rule is sometimes inconsistent as there are others (exceptional cases) with ending in 'e', 'o', 'i', 'u' as the root forms. Same goes with plural formation in Hausa nouns, endings in 'a' are typically considered plurals. For example, the word "mata" (women) from "mace" (woman) and "yara" (boys) from "yaro" (boy), etc. But in some cases, endings in 'u' are plurals, for instance, the word "kujeru" (chairs) from "kujera" (chair). Therefore, there are also singular words ending in 'a'.

Hence, the stemming errors in this system are as a result of these inconsistencies. To overcome the mistakes, find all these exceptional cases and employ them in the system.

V. Conclusion

It is possible to develop Hausa morphological stemmer and improve it over time despite the scarcity of linguistic resources. The proposed system can efficiently stem Hausa word with accuracy of 73.8%. The system is quite good even though it lacks abundant exceptional cases. The results showed that Hausa stemming should not depend on a dictionary but rather on exempting some particular group of words from the process. Over-stemming and under-stemming with 26.2% are, often, due to this unavailability. Provision of such exceptions will enhance the efficiency of the system.

Our future work includes the provision of more exceptional cases (Hausa proper nouns inclusive) as well as additional rules to help in improving the system. Also, future work will involve utilizing part of speech tagging. Part of speech tagging will be conducive in further dividing the Hausa stemming process into several modules such as noun and verb modules stemming.

Acknowledgements

The authors would like to extend appreciation of the support given by University Sultan Zainal Abidin (UniSZA) and the Kano State Government. Also, our gratitude to Dr. Tijjani Shehu (a lecturer, Department of Linguistics, Bayero University, Kano) for his tireless support in examining the results.

References

[1]. Lovins, Julie B. Development of a stemming algorithm. MIT Information Processing Group, Electronic Systems Laboratory, 1968.

[2]. Frakes, W. B. "Introduction to information storage and retrieval systems." Space 14 (1992): 10.

[3]. Abdullah, Syed, Engku Fadzli Hasan, Syarilla Ahmad Saany, Hasni Hassan, Mohd Satar, and Siti Dhalila. "Simple Rules Malay Stemmer." In The International Conference on Informatics and Applications (ICIA2012), pp. 28-35. The Society of Digital Information and Wireless Communication, 2012.

[4]. Pande, B. P., Pawan Tamta, and H. S. Dhami. "Generation, Implementation and Appraisal of an N-gram based Stemming Algorithm." arXiv preprint arXiv:1312.4824 (2013).

[5]. Dianati, Mohammad Hassan, Mohammad Hadi Sadreddini, Amir Hossein Rasekh, Seyed Mostafa Fakhrahmad, and Hossein Taghi-Zadeh. "Words Stemming Based on Structural and Semantic Similarity." Computer Engineering and Applications Journal 3, no. 2 (2014): 89-99.

[6]. Smirnova, Mirra Aleksandrovna. The Hausa language: a descriptive grammar. Routledge & Kegan Paul, 1982.

[7]. Newman, Paul. The Hausa language: An encyclopedic reference grammar. Yale University Press, 2000.

[8]. Hafer, Margaret A., and Stephen F. Weiss. "Word segmentation by letter successor varieties." Information storage and retrieval 10, no. 11 (1974): 371-385.

[9]. Adamson, George W., and Jillian Boreham. "The use of an association measure based on character structure to identify semantically related pairs of words and document titles." Information storage and retrieval 10.7 (1974): 253-260.

[10]. Mcnamee, Paul, and James Mayfield. "Character n-gram tokenization for European language text retrieval." Information retrieval 7.1-2 (2004): 73-97.

[11]. Frakes, William B. "Term conflation for information retrieval." Proceedings of the 7th annual international ACM SIGIR conference on Research and development in information retrieval. British Computer Society, 1984.

[12]. Porter, Martin F. "An algorithm for suffix stripping." Program 14, no. 3 (1980): 130-137.

[13]. Lessa, Lemma. "Development of stemming algorithm for wolaytta text.", master's thesis, Department of Information Science, Faculty of Informatics, Addis Ababa University, Ethiopia (2003).

[14]. Khan, Sajjad Ahmad, Waqas Anwar, Usama Ijaz Bajwa, and Xuan Wang. "A Light Weight Stemmer for Urdu Language: A Scarce Resourced Language." In 24th International Conference on Computational Linguistics, p. 69. 2012.

[15]. Abdoulaye, Mahamane L. "Aspects of Hausa morphosyntax in role and reference grammar." PhD diss., State University of New York at Buffalo, 1992.

[16]. Al-Hassan, Bello SY. "Transfixation in Hausa: A Hypothetical Analysis." Studies of the Department of African Languages and Cultures 45 (2011).