LEXICON AND RULE-BASED LEMMATIZATION APPROACH FOR SOMALI LANGUAGE

ABSTRACT

Lemmatization is a Natural Language Processing (NLP) technique used to normalize text by changing morphological derivations of words to their root forms. It is used as a core pre-processing step in many NLP tasks including text indexing, information retrieval, and machine learning for NLP, among others. This paper pioneers the development of text lemmatization for Somali, a low-resource language with very limited or no prior effective adoption of NLP methods and/or datasets. Especially, we develop a lexicon and rule-based lemmatizer for Somali text, which is a starting point for a full-fledged Somali lemmatiation system to be used for various NLP tasks. With consideration of the language morphological rules, we have manually crafted an initial lexicon of 1127 tokens (root words) with 6240 candidate lemmas. We have tested the algorithm on fifty-five documents of various lengths including general news articles, social media posts, and text messages. Our initial results demonstrated that the algorithm achieves an accuracy of 64.4% for relatively long documents, and about 68.6% for short texts.

1 INTRODUCTION

An estimated population of over 30 million, mainly in Somalia, Djibouti, Kenya, Ethiopia, the UK, the USA, and Europe speaks the Somali language. Besides, we all know that there is no way that one can understand and access information better than using one's own language, however, there are significant digital limitations associated with the Somali language preventing it from fully participating in today's digital world. For example, Somali is a low-resource language, which means that there are very limited or no datasets available for AI research and applications like translation, transcription, classification, language modeling, and many others. This puts the Somali speaking people at disadvantage and limits them from utilizing such essential AI and NLP technologies. As such, the initial research in the paper will be a starting point in the study of lemmatization in the Somali language and the construction of a lexicon dedicated to this task.

Lemmatization is an NLP method of transforming words of the same root to their base shape, and there are different methods that can be used to find the root words, for example using rules or a dictionary-like look-up method (lemma), or a combination of methods. Different NLP methods and lemmatization techniques have been proposed and datasets created for many developed main languages, e.g. English, French, etc, but such resources are not available for many "low-resource" languages including the Somali language and this is where the importance of this study comes in.

In the present paper, we study the problem of lemmatization in the Somali language to develop an approach for normalizing words derived from the same roots. The main objective of lemmatizer is to extract meaningful root words from inflectional words by applying look-up and rules-based methods. Especially, this study mainly focuses on the lemmatization of the "MAXAA TIRI", which is the main and written dialect of the Somali Language. In other words, we built two ways to find the root words: a lexicon-based or dictionary (i.e., searching words in the dictionary) and rule-based (looking at the beginning of the term and determining the root of the given word). For the lexicon-based, we constructed an initial dictionary built on the morphological association of words enriched. Finally, the combined lexicon and rule-based method are applied to test dataset of over 50 documents achieving an average lemmatization accuracy close to 70%. This performance shows that there is a potential for developing a full-fledged Somali Lemmatizer by scaling the lexicon and consolidating the rules.

The contributions of this work are three-fold:

1. First, we have developed an initial and Somali lexicon for word lemmatization with the consideration Somali word morphological rules.

2. Second, we have designed a Somali word lemmatization algorithm built on the

3. Third, we have tested the algorithm on fifty five documents of various lengths including general news articles, social media posts, and Whatup messages to estimate the performance of the proposed lemmatization algorithm.

2 RELATED WORK

Lemmatization is a well-studied research problem of well-resourced languages, e.g. English, French, etc, but remains an open research problem for low resourced languages such the Somali [8, 3, 5]. Besides, lexicon and rule based approaches are some of the most common methods used for lemmatization particularly for languages with limited resources. For examples, authors in [7] study the development of lemmatization algorithms based on if-then-rules for Slovene. Similarly, Prathibha and Padma [9] put forward a rule based lemmatization algorithm for Kannada language by combining linguistic rules and meaningful root words. On the other hand, the work by Eger et. al. [2] conducted a comparative study of lexicon-based approaches for Latin and they found that lexicon-based lemmatization can achieve high accuracy if the considerable dataset is in place.

In addition, very few studies have been done about the application of NLP methods in the Somalia language and/or the construction of related datasets. In one of the very few related studies, Mohammed [6] investigates the problem of part-of-speech (PoS) tagging for Somali language using statistical and machine-learning approaches. The auhots claimed that their Somali POS tagger achieved a high accuracy of 87.51% on tenfold cross-validation experiments. Furthermore, the work Abdillahi of et. al. [1] presents a study on the automatic transcription of the Somali Language. Their work constructs a corpus from audio-speech of about 10 hours achieving about a 21-word error rate (WER).

To the best of our knowledge, there have not been any previous works on the subject of lemmatization for the Somali language, which makes this work the first of its kind for normalizing text written in Somali, paving the way for further research on the creation of NLP resources and techniques for the Somali Language.

3 PROPOSED LEMMATIZATION APPROACH

Figure 1 provides a high-level illustration of the proposed Somali Language lemmatizer. As shown, constructing the lexicon is the first step of the method in which rules for pairing root and derived words were developed before commencing the actual compilation of the lexicon. The lexicon is based on the root word being made into a key, and the words following it are made into values. Although on the other hand, the root word can only be a verb if the word is a command. Because of the limited size of the lexicon created in this initial work, we have also developed simple rules for handling words that are not found in the lexicon to improve the lemmatizer accuracy. The numbers between the lexicon/written rules and the lemmatization method in the figure indicate the lemmatizer's order to normalize text. In other words, it first attempts to find the lemma from the lexicon and applies the written rules if the root word cannot be found in the dictionary.

The lemmatization proposed in this work is particularly for the main Somali dialect "MAXAA TIRI". To perform the lemmatization, we first read the lexicon or dictionary we have prepared. The loaded data is then converted to a JSON format to enable us to further process the lexicon and retrieve the root words. Before extracting the lemmas, we applied a bit of data such as removing the stopwords the target of the lemmatizer is keywords. Finally, we look for root words in the lexicon in the JSON format. If the base form is found, it will be returned. If it is not found, another rule-based algorithm is looked for in the same way; if it is found, the words of these two algorithms will be recovered and will be counted.

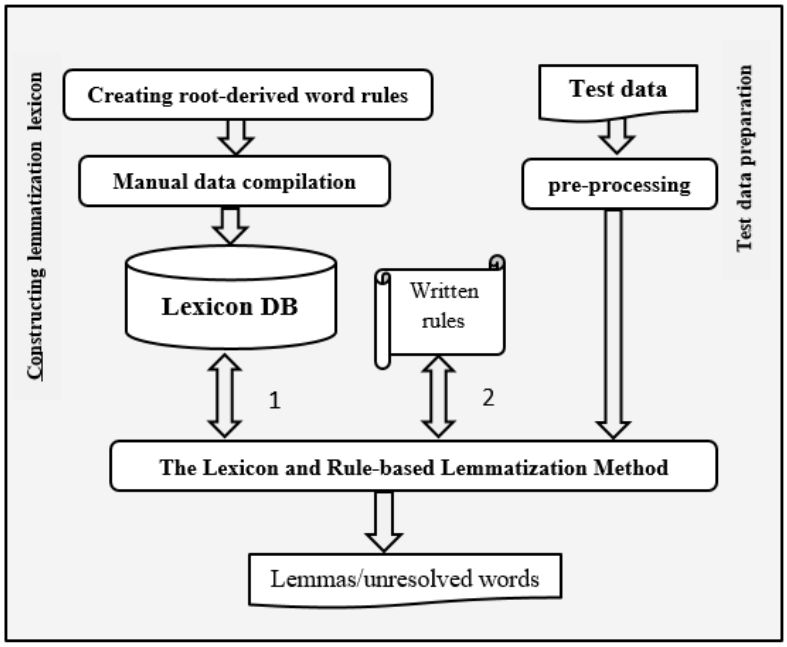

Figure 1: Proposed lexicon and rule based lemmatizer

3.1 LEXICON

Using a variety of sources, we have collected the most commonly used words in the langague randomly and without regard to a specific domain, e.g. politics, education, health, and so on. Examples taken from the collected works include “baarlamaan” (a political term), “jirran” (medical term) and “macallin” (an academic or education-related term). The dataset is collected in a dictionary we designed based on simplicity in terms of reading, adding, formatting, and the way we verify the data rather than the sources. The lexicon consisted of a total of 7367 tokens including 1127 root words and 6246 derivatives. In the process of the lexcion construction, we separated the words into verbs and nouns. Based on the langauge morphology, most verbs of the Somalia can be changed to root by returning them to the command verb, for example, ”cab”, ”cun”,”laq” and so on. Regarding the nouns, the origin of the noun is singular for example ”badar”,”sonkor”,”caano”, etc. The dictionary is structured as shown in the figure below. Each word may have different forms and also may have different meanings apart from the root word, for example, a lexicon entry (root word and its derivatives) is like: /

Example 1 for the root verb ”cab” (eat): cab:[”caba”,”cabay”,”cabeen”,”caba”,”cab”]

Example 2 for the root noun ”bariis” (rice): bariis:[”bariis”,”bariiska”]

You may notice that word derivations are reflected with changes in the suffixes, or prefixes. We have built the lexicon in the form of a key and values as illustrated in the above examples. This structure helps in the search for the root word, i.e., if the key is found in the entry it can be returned while reducing the time complexity of the algorithm.

3.2 RULES

To expand the lemmatized vocabularies and look at the simplest ways to store words, we have adopted a rule-based form in which the word is derived from the beginning. However, there may be similar words as a pseudo-general formula in the beginning. If the word starts with a specified word, then the origin of the word specified and returned, etc. Algorithm 1 summarizes the base word identification process.

The pseudocode in Algorithm 1 is complementary to Figure 1 and it shows how the root word look-up and application of the rules work for this algorithm. First, the sentences are given to be processed

Algorithm 1: get root words from a list of rules:

```
(i) INPUT: Load each Sentence;
(iii) Perform word Tokenization
(iv) Remove stop words and duplicate words;
(v) for each word do
    for key and value in lexicon do
        search and find specific word;
        if word inside of the lexicon then
            return specified root word which is the key from the lexicon
            add statistics results;
        end
        assume that the word not found in the lexicon next step is to search
        rules;
        if word starts with the way we specified then
            return specified root word
            add statistics results;
        end
        otherwise do
            Error: unresolved word
        end
    end
end
```

and tokenized. If the word normalization goes to the rules level and rules matching a word are found, the process will return the root word that we defined before, otherwise, it will be recognized as an unresolved word.

4 EXPERIMENTAL SETUP

As mentioned previously, this study is done for Somali texts written in the "MAXAA TIRI" (aka Somali-Somali). Empirically speaking, to find the root for a given word in our dataset, a simple search is applied and if the word is not found in the dataset, it will be searched in the list of rules. A user of this algorithm will supply a text and will pass it to the algorithm, the text will be processed and broken down word by word so that each word is looked for separately and the list in the dataset for the word found will be returned.

The primary metric utilized to assess the effectiveness of the suggested lemmatizer is accuracy. The measure is calculated as the ratio of successfully lemmatized words to all the different input-inflected words. The following equation illustrates the suggested lemmatizer's mathematical precision [4], but in our study, there is a slight difference as shown below.

$$Accurracy(\%) = \frac{Totalwordsresolved}{documentlength}$$

4.1 RESULTS AND DISCUSSION

In this section, we present sets of tests to evaluate the effectiveness of the lemmatization method built on the lexicon and rule-based approach. Indeed, in a low-resource language preparing lemmatization is not easy, not only due to the limitation of the required morphologicla data but also due to the difficulty of objectively judging the quality of lemmas. In practice, most dictionary keys and values are used as baseline criteria for evaluating methods. If the algorithm corresponds to the values that were manually created, the key-based solution works well. However, in practice, datasets often are manually created and the searched data, i.e., what needs to be normalized, comes from the user. In our initial evaluation, we collected test documents of different sizes tested the documents, and preprocessed the data by removing stop words, duplicates, and all special characters. In other words, the lexicon we have created has reached up to 1127 words in terms of roots, but each word has many derivational words. Sometimes, it is possible for each word to generate more than 20 derivative words. From the initially constructed data, the derivative words reached 13524.

The document in Figure 2 is one of the tests we have done, although the length of this one is relatively small. There are other documents that are larger than this in the test set. When we applied the lemmatization algorithm to this example document, we obtained the results presented in Table 1.

Waa Tartanka Hurdada oo Lagu Qabta Koonfurta Kuuriya Hal Saac Hadaad Hurudo Waxaa Lagu Siinayaa 1,400$ Qofba Qofka uu Ka Hurdo Badanyahayna Waa uu kasii Lacag Badanyahay. Waa Mid Kamid ah Tartamada ugu Yaabka Badan aduunka Adiga Imisa/Meeqo saac baad Hurdi Kartaa? Abdifatah Haji.

Figure 2 :Short Illustrative document

Source	Result
origin docs size in words	40
stop words(non-unique)	7
special characters	7
unresolved words	5
unresolved words	23
percent found	82%
words from lexicon	21
words from rule-based	2

Table 1: statistical results.

As you can see, the document size was 40 words, and 7 stop words were identified which were part of the total document word count. It also contains 5 special characters in the same way that stop words are counted. The percentage of words that were resolved or the words that were found seem to be high as this reaching 82.14%. This may be due to the fact that this is a short document. As shown in the statistics, the lemmatizer first applies the look-up, the words found are recorded. It then moves to the next level of using the rules. Any words not lemmatized at this stage are termed unresolved to be recognized. Overall and part of this illustrative document, we have tested the algorithm on 55 documents of various lengths including general news articles, social media posts, and Whatup messages. The results we found for these different document categories are summarized in Table 2. As shown, the results demonstrate that the algorithm achieves a relatively good accuracy of 64.4% for moderately long documents (e.g. general news), and about 68.6% for short texts such as social media and text messages. This initial result is promising for a language with no previous or very limited NLP resources. From these results, it can be thought that the lexicon and the rule-based slightly work well for social text messages. From the tests, we have noticed that most word normalization is achieved at the dictionary look-up level, specifically for short messages.

Type	# docs	Avg Doc Len	Avg Acc
General News	25	193.84	64.37%
Social Media	25	57.76	68.08%
Text Message	5	25.2	68.59%

Table 2: documents in length evaluation.

5 Conclusion and future work

This work is the first of its kind to develop a text lemmatization for the Somali language. It proposed a lexicon and rule-based lemmatizer for Somali text as a starting point for a full-fledged Somali lemmatization system NLP tasks. Considering the morphological language rules, we have manually crafted an initial lexicon of 1127 root words and 6240 derived words to serve as a dictionary look-up database for the Somali lemmatization task. Specifically, the study focuses on the lemmatization of the Somali "MAXAA TIRI" (Main Somali dialect). Furthermore, we have built two ways to find the root words: a lexicon-based or dictionary look-up method. The lexicon is achieved by searching in the dictionary, and the rule-based is to look at the beginning of the term and then determine the root of the given word. Finally, we did initial tests and found some promising results of over 68% lemmatization accuracy for short documents. This meant that the study has reached a reasonable level of accuracy, paving the way for further research in enabling the Somali with NLP techniques.

As our research is still in progress, we plan to further improve the proposed lemmatization by increasing the size of the lexicon database and expanding the applied rules. We will also investigate

the possibility of devising a methodology for building the lemmatization lexicon automatically from a collected Somali corpus text.

ACKNOWLEDGMENTS

First, we thank the almighty Allah for enabling us to do research in this exciting field. Secondly, We would like to express our gratitude and thankfulness to the Somali Academy Of Science And Arts for spending a lot of time with us and helping us check the accuracy of the Somali language words. Finally, we thank everyone who volunteered to contribute to this work.

REFERENCES

[1] Nimaan Abdillahi, Nocera Pascal, and Bonastre Jean-François. "Towards automatic transcription of somali language". In: *les actes de Language Resource and Evaluation Conference (LREC)*. 2006, pp. 131–134.

[2] Steffen Eger, Tim vor der Brück, and Alexander Mehler. "Lexicon-assisted tagging and lemmatization in Latin: A comparison of six taggers and two lemmatization methods". In: *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*. 2015, pp. 105–113.

[3] Michael A Hedderich et al. "A survey on recent approaches for natural language processing in low-resource scenarios". In: *arXiv preprint arXiv:2010.12309* (2020).

[4] Julie Beth Lovins. "Development of a stemming algorithm." In: *Mech. Transl. Comput. Linguistics* 11.1-2 (1968), pp. 22–31.

[5] Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. "Improving lemmatization of non-standard languages with joint learning". In: *arXiv preprint arXiv:1903.06939* (2019).

[6] Siraj Mohammed. "Using machine learning to build POS tagger for under-resourced language: the case of Somali". In: *International Journal of Information Technology* 12.3 (2020), pp. 717–729.

[7] Joël Plisson, Nada Lavrac, Dunja Mladenic, et al. "A rule based approach to word lemmatization". In: *Proceedings of IS*. Vol. 3. 2004, pp. 83–86.

[8] Martin F Porter. "An algorithm for suffix stripping". In: *Program* 14.3 (1980), pp. 130–137.

[9] RJ Prathibha and MC Padma. "Design of rule based lemmatizer for kannada inflectional words". In: *2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)*. IEEE. 2015, pp. 264–269.