Construction of Morpheme-Based Amharic Stopword List for Information Retrieval System

Tilahun Yeshambel[1(✉)], Josiane Mothe[2], and Yaregal Assabie[3]

[1] IT PhD Program, Addis Ababa University, Addis Ababa, Ethiopia
tilahun.yeshambel@uog.edu.et
[2] INSPE, Univ. de Toulouse, IRIT, UMR5505 CNRS, Toulouse, France
josiane.mothe@irit.fr
[3] Department of Computer Science, Addis Ababa University, Addis Ababa, Ethiopia
yaregal.assabie@aau.edu.et

Abstract. One of the major forms of pre-processing in information retrieval and many other text processing applications is filtering out stopwords. They are ignored by many retrieval systems during indexing and retrieval in order to enhance retrieval effectiveness and efficiency. The aim of this paper is to present the construction of morpheme-based Amharic stopwords and investigate their effect on information retrieval tasks. The stopword list is constructed based on the semantics of Amharic words and corpus statistics: frequency, mean, variance, and entropy parameters. The stopword list is evaluated using Lemur on Amharic information retrieval test collection. Removal of stopwords has shown significant impact on retrieval effectiveness, size of index and term weighting of non-stopwords. On the other hand, their presence in index and query negatively affects the retrieval effectiveness of Amharic retrieval system. The average precisions of retrieving with and without stopwords using language modeling on root-based approach are 0.24 and 0.70, respectively.

Keywords: Morphological analysis · Corpus statistics · Semantics · Complex-language · Amharic · Stopword

1 Introduction

Stopword identification is one of the important tasks of text processing applications such as text classification, Information Retrieval (IR), text summarization, etc. [1]. In many languages, stopwords are functional and general words with low discriminatory power to differentiate between text documents, and affect efficiency [2]. They include articles, conjunctions, personal pronouns, prepositions, etc. These are frequently occurring words in a natural language and evenly distributed words across documents in corpus. They make up large portion of the text. They are considered as unimportant in many applications and thus, many text preprocessing applications remove them before processing documents and queries to improve system performance, save memory space and processing time [1, 3]. The two commonly used stopword removal techniques are the

M. A. Delele et al. (Eds.): ICAST 2020, LNICST 384, pp. 484–498, 2021.
https://doi.org/10.1007/978-3-030-80621-7_35

use of stopword list and inverse document frequency (IDF) value. For example, natural language toolkit (NLTK) has lists of stopwords for 16 different languages[1], and [4] removes stopwords using their IDF values.

IR is the process of finding relevant documents for users' query. Stopwords are ignored by many IR systems during indexing and query processing [2, 5]. In IR, stopword list contains frequent words that are ineffective to distinguish one document from others. Elimination of these words from the index reduces the space requirement and increases retrieval effectiveness in different languages. Depending on the complexity of the language, they can be removed before or after text processing such as stemming. The retrieval systems such as Okapi [6], Terrier [7], Lemur[2] and Lucene4IR[3] use stopword list to improve retrieval effectiveness.

Stopword lists can be constructed either manually or automatically. Manual construction involves the analysis of the semantics of words in a given language whereas automatic construction of stopword lists is using statistics information from large corpus. For example, Feng *et al.* [3] built a Chinese stopword list using TREC (Text Retrieval Conference) Chinese corpora. The list was constructed automatically based on the combination of the mean of probability, variance of statistical model and entropy of information model. On the other hand, Hindi text stopword list was constructed based on term weighting and relevance of words with respect to corpus [8]. The final generic stopword list was compiled based on the aggregation of term weight and entropy value. The Persian stopword list was generated based on terms' frequency, normalized IDF and information model [9]. Words with high-frequency, low IDF value, and with high entropy values are considered as stopwords, which belong to adverbs, prepositions, interjections and auxiliary parts-of-speech. As reported in this research, removal of stopwords minimizes the number of index terms by 27%. A standard stopword list for Malay language was built using word's frequency, variance, and entropy from a corpus that has 7,363,578 tokens [5]. Words with highest frequency, variance and entropy values were considered as stopwords. The intersection of the top n words from term frequency, variance and entropy were used to generate the final stopword list that contains 339 words. There are also other stopword lists for various languages such as Chinese, English, French, German, Arabic, Portuguese and Spanish languages[4].

Few attempts were made to create Amharic stopwords [10]. Mindaye et al . [11] created manually a stopword list which contains 77 words. However, this list has problems. First, the list contains small number of stopwords. From the actual number of stopwords available in the language, this list ignores many stopwords. Second, the list contains variants of a stopword rather than basic stems. Therefore, in one side it is difficult to list all variants of each stopword. On the other hand, using all variants of stopwords increases the length of the list. This creates unnecessary computational time specially on online processing of texts such as query processing. Samuel and Bjorn [12] create news specific stopword list which contains 745 words automatically. Like Mindaye et al. [11], this list contains variants of stopwords. Since this list is domain specific, it

[1] https://www.geeksforgeeks.org/removing-stop-words-nltk-python/.
[2] http://www.lemurproject.org/.
[3] https://github.com/lucene4ir/lucene4ir.
[4] http://www.ranks.nl/stopwords/, http://github.com/Alir3z4/stop-words.

could not be applicable in other domains. Alemayehu and Willett create corpus based stopword list which contains 148 stems [13]. In addition to the non-content bearing words, numerals are included in this list. The list contains the basic stems and citation forms of stopwords. However, there are more than one stems for variants of some stopwords, {ነበር, ነባር}, {ደግም, ዳግም}, {ሁን, ሆን}, e.g. etc. Generally, it can be observed that the construction of stopwords for Amharic IR system attempted so far has not followed the scientific methods used to create a list of stopwords and cannot be used as a generic resource for Amharic IR. As a result, some research and development works on Amharic IR do not remove stopwords [14–17] while others remove based on non-standard stopword list created manually [11, 18] or automatically [4]. The manually created stopword lists are compiled for a specific purpose and are randomly handpicked without proper consideration of the techniques employed to identify stopwords. Furthermore, the lists include only few stopwords as samples. On the other hand, the automatically constructed stopword list considered only IDF values and does not take into account stopword distribution in Amharic texts and the characteristics of the language.

The morphological characteristics of languages play a significant role in the identification of stopwords. Words generated from the same root in morphologically simple languages have one common stem. For such languages, a stemmer can conflate variants to one common form and thus, the frequency information about each term can be computed after applying stemming. As a result, terms which are distributed across many documents in the corpus can be identified by their IDF values. In such cases, IR systems developed based on term frequency are performing well. However, morphologically complex languages like Amharic have multiple stems for words generated from the same root. For instance, words such as ሰበረ */səbərə* 'he broke'/, ተሰባሪ/*təsəbari* 'that can be broken'/, and አሳበረ */ʔəsabərə* 'he helped to break'/ have the basic stems ሰበር- */səbər-/*, ሰባር- */səbar-/* and ሳበር- */sabər-/,* respectively. Hence, the use of stems provides distorted frequency since each stem is considered as different. Thus, Amharic stems need one more reduction analysis to extract their common form which is the root. In the aforementioned example, the three stems have one common root represented as ስ-ብ-ር */s-b-r/*. Hence, collection statistics can be computed accurately based on root forms. This calls for the application of morphological analysis before the identification of stopwords in Amharic IR. However, to our best knowledge, there is no systematically constructed Amharic stopword list so far that considers term statistics and morphological characteristics of the language. Therefore, the aim of this paper is to generate Amharic stopword list using frequency, mean, variance and entropy values of root forms of words.

The rest of this paper is organized as follows. Section 2 briefly describes the characteristics of Amharic language. Section 3 presents the construction of Amharic stopword list. Experimental results and discussion are presented in Sect. 4. Section 5 highlights the effect of the constructed stopwords in Amharic IR. Finally, conclusion and future research directions are forwarded in Sect. 6.

2 Amharic Language

Amharic is the working language of the government of Ethiopia. It belongs to Semitic languages families. The language uses Ethiopic script for writing, which has 33 basic

characters where each of them has 7 different forms representing consonant-vowel combination [4]. On top of basic characters there are labialized characters such as ሏ/*lwa*/, ሟ /*mwa*/, ሷ /*swa*/, ቋ /*qwa*/, ሯ /*rwa*/, etc. Their structure is consonant-vowel-vowel combinations [4, 19]. In addition, the script has its own punctuations and numbers. Its rich literary heritage has endowed the language with huge written resources.

Amharic is morphologically rich and complex language. The word formation process undergoes complex inflectional and derivational process [20]. Words can be formed directly from their roots by inserting vowels between radicals, from stems by attaching affixes, reduplication of one of the character of the stem or word itself, or compounding. Inflectional and derivational words change their forms for different purposes such as nouns, adjectives, adverbs, and verbs formation. Thousands of surface words can be generated from an Amharic root and its stems by changing the shape of characters in a stem or root, and by attaching affixes on stems [20]. In Amharic, verbal stems are derived from root's consonants by interdigitating vowel patterns. There are different templates/stem structures for the formation of words in various forms such as perfective, imperfective, jussive, imperative, etc. [21]. Amharic has many morphemes that play significant roles in morphology and syntax. Morphemes appear as affixes that have their own functions carrying different types of syntactic and semantic information. Amharic affixes are classified as prefix, suffix, infix, and circumfix, which might be added at the beginning, end, inside, or both at the beginning and end of the stems, respectively. They are attached to the base forms to mark for gender, number, case, person or others, and give additional functions to the roots or stems of words. More than one morpheme might exist on a given word. According to [22], an Amharic word might take up to four prefixes and five suffixes. For example, the word የሚያስመጣላቸውን /*jəmijasmət'alatʃəwin*/ is made from four prefixes (የ /*jə* preposition/, ም/*m* nominalizer/, ይ /*ji* third person singular subject marker for imperfect verb/, and አስ /*ʔəs* causative/); imperfect verb (መጥ /*mət'/*); and four suffixes (ኣ /*ʔa* third person singular subject marker for imperfect verb/, ልlli beneficative/, ኣቸው/*ʔatʃəw* object marker for third person plural/, ን /*n* accusative/). Due to such level of complexity, the morphological structure of Amharic is an important issue in the development of natural language processing applications.

3 Construction of Amharic Stopword List

Stopwords are generally useful for providing a good structure for a text. However, they contribute little meaning to the content. They can be identified using different techniques such as dictionary-based and automatic approaches. Dictionary-based approaches are inefficient and very expensive as stopwords are selected manually. On the other hand, automatic methods construct a list of stopwords based on statistical information such as term frequency, IDF, and variance from a large corpus. The characteristics of languages may play significant role in selecting the techniques employed for the construction of stopword lists. Considering the characteristics of Amharic, we built domain independent stopword list by combining two methods: *semantic-based* and *corpus-based*. Semantic-based identification of stopwords is carried out by analyzing the nature or semantics of words in the language whereas corpus-based method applied by considering the statistical information of words in a corpus.

3.1 Semantic-Based Identification of Stopwords

Based on the semantics of words, Amharic words can be classified into two as main and sub-word classes. The main word classes include verbs, nouns, adjectives and adverbs whereas sub-words include exclamation, conjunction, preposition, etc. Main word classes provide meaning by their own. On the other hand, sub-words are functional words used mainly for the formation of phrases, sentences and paragraphs. They occur commonly in many descriptions of different events. However, they contribute little to the description of the topics covered by the text, and they have structural function rather than meaning. According to [23], Amharic sub-word classes are characterized by lack of meaning by their own, inability to undergo morphological derivation and inflection, lack of morphemes for various parameters, and they have small word size. Amharic stopwords which satisfy these characteristics are collected from an Amharic book "Yamarigna Sewasiw" [23]. It includes words such as ወደ/*wədə* 'towards'/, እንደ /*ʔində* 'like'/, ስለ /*sɨlə* 'about'/, እስከ /*ʔɨskə* 'up to'/, ወዘተ /*wəzətə* 'and so on' /, ጎሽ /*goʃ* 'bravo'/, ኡ /*ʔu* 'oh'/, ዋ /*wa* 'warning'/, ይልቅ /*jɨlɨk* 'instead'/, etc.

3.2 Corpus-Based Identification of Stopwords

Corpus-based identification of Amharic stopwords is made by analyzing the frequency, mean, variance and entropy of words in Amharic text corpus. Non-content-bearing words in Amharic are usually used to structure sentences and paragraphs. They are used for keeping the coherence of a text rather than describing the subject matter. Non-content-bearing words are also used as morphemes in Amharic word formation. Accordingly, morphological analysis needs to be done before identification of stopwords. Thus, statistics of terms is computed on morphologically analyzed text corpus. The process of corpus-based stopword identification is presented in Fig. 1.

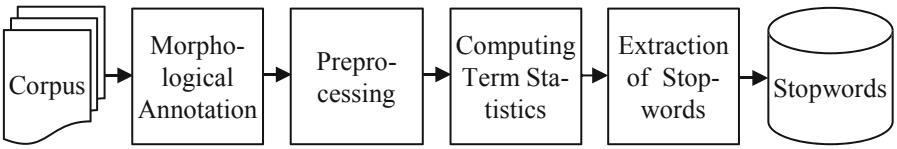


Fig. 1. The process of corpus-based Amharic stopword identification

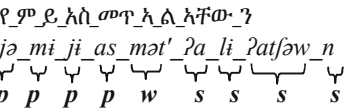
Corpus Collection: Identification of many non-content-bearing Amharic words is carried out using statistics of terms computed from a representative corpus. To this effect, we collect documents from Amharic Wikipedia[5], religious books, blogs[6], and news sources[7] to create a corpus. Thus, the corpus has 5,737 documents which contain 64,637 sentences and 1,315,371 words. The corpus is built for the general purpose of Amharic IR test collection as well.

[5] https://am.wikipedia.org.
[6] http://www.danielkibret.com/.
[7] http://www.waltainfo.com/.

Morphological Annotation: All words in the corpus are annotated with morphological segments. The annotations segment surface words into their constituent parts called morphemes based on stems and root forms. The morphemes can be preposition, stem, root, person marker, tense marker, conjunction, plural marker, etc. As the corpus is built for general purpose, all morphemes in each word are tagged with morphological information. The general structure of a morphologically annotated word ***W*** is:

$$[\boldsymbol{p}\_]^{*}\boldsymbol{w}[\_\boldsymbol{s}]^{*}$$

where p is a prefix morpheme, "_" is a morphological segment marker, *w* is the root or stem of ***W***, *s* is a suffix morpheme, […] denotes optionality, and * denotes the possibility of multiple occurrence. For example, the word የሚያስመጣላቸውን/*jəmijasməṭ'alatʃəwn*/ is annotated as follows.



Preprocessing: Preprocessing of the annotated corpus involves character normalization and removal punctuation marks and. A major task of preprocessing task in Amharic corpus is normalization. The process is required due to the fact that some characters have similar phonemes but different graphemes. Base characters having such property are {ሀ /*hə*/, ሐ /*hə*/, ኀ /*hə*/ and ኸ /*hə*/}, {ሠ /*sə*/ and ሰ /*sə*/}, {ጸ /*ts'ə*/ and ፀ/*ts'ə*/}, and {አ /*ʔə*/ and ዐ /*ʔə*/}. Thus, ስለጸሀፊዋ can also be written as ሥለጸሀፊዋ,ስለጸሐፊዋ, ሥለጸሐፊዋ, ሥለጸኀፊዋ,ስለጸኀፊዋ, ሥለጸሐፊዋ, etc. although some of them rarely appear in a text. As there is no standard established for character normalization, we used the rules shown in Table 1 to normalize characters. Furthermore, due to similarity in phonemes, the fourth orders ሃ /*ha*/, ሓ /*ha*/, ኃ /*ha*/ and ኻ /*ha*/ are normalized to ሀ/*hə*/ whereas ኣ /*ʔa*/ and ዓ /*ʔa*/ are normalized to አ /*ʔə*/.

Table 1. Character normalization

Base character	Normalized form	Example	
		Original word	Normalized word
ሐ	ሀ	ሓረግ	ሀረግ
ኀ		ኀይል	ሀይል
ኸ		ኽሰት	ሁሰት
ሠ	ሰ	ሢሣይ	ሲሳይ
ዐ	አ	ዓላማ	አላማ
ፀ	ጸ	ፅሑፍ	ጽሁፍ

Computing Term Statistics: In the development of IR systems, term statistics has been widely used to identify stopwords. However, the notion of terms depends on the characteristics of languages. For morphologically simple languages, stems can be considered

as terms. However, this is not the case with morphologically complex languages like Amharic. We hypothesize that morphemes used to form Amharic words could be used as a basis for computing term statistics. Thus, in this work, we consider morphemes as terms. Accordingly, for the entire corpus that we collected, we compute morphemes frequency, mean, variance and entropy as described below.

Document and Collection Frequency: The document frequency of a morpheme indicates the number of documents the morpheme exists whereas collection frequency is the total morpheme frequency throughout the corpus. In this case, all morphemes from all documents are ranked according to their document frequency and collection frequency. Then, a threshold value was set to determine stopwords from ranked morpheme. Furthermore, morphemes that are evenly distributed throughout the collection and satisfy the threshold value are considered as stopwords [24]. Document frequency $\boldsymbol{df}$ computed as:

$$\boldsymbol{df}(\boldsymbol{M_i}) = \sum_{\mathbf{i}=1}^{\mathbf{N}} \mathbf{morpheme\_\ status}(\boldsymbol{D_i}) \tag{1}$$

where $\boldsymbol{M_i}$ is the $\boldsymbol{i}^{\text{th}}$ morpheme in the corpus, $\boldsymbol{D_i}$ is the $\boldsymbol{i}^{\text{th}}$ document in the corpus, $\boldsymbol{N}$ is the total number of morphemes in the collection. If a morpheme appears in a given document, its status is 1 otherwise 0. Collection frequency $\boldsymbol{cf}$ is computed as:

$$\boldsymbol{cf}(\boldsymbol{M_i}) = \sum_{\boldsymbol{i}=1}^{\boldsymbol{N}} \boldsymbol{MFD}_i \tag{2}$$

where $\boldsymbol{MFD_i}$ is the morpheme frequency in each document, $\boldsymbol{N}$ is total number of documents in the corpus. In this paper, a morpheme frequency is un-normalized total number of times a morpheme occurs in a document.

Mean: This is another way to measure the overall distribution of morphemes in the whole corpus. The mean probability $\boldsymbol{mp}$ of each unique morpheme in the whole corpus is computed as:

$$\boldsymbol{p}(\boldsymbol{M}_\mathrm{i}) = \frac{\boldsymbol{MF}}{\boldsymbol{TM}} \tag{3}$$

where $\boldsymbol{p}(\boldsymbol{M_i})$ is morpheme probability, $\boldsymbol{MF}$ is morpheme frequency in each document and $\boldsymbol{TM}$ is the total number of morphemes in the document. Then, the mean probability of each morpheme, $\boldsymbol{mp}\ (\boldsymbol{M_i})$, in all documents is computed using as:

$$\boldsymbol{mp}(\boldsymbol{M}_i) = \frac{\sum_{\mathbf{i}=1}^{\mathbf{N}} \mathbf{p}(\mathbf{M_i})}{\mathbf{N}} \tag{4}$$

where $\boldsymbol{N}$ is total number of documents.

Variance: This is the other way to check the distribution of morphemes throughout the documents in the corpus. Variance $\boldsymbol{v}$ is computed as:

$$\boldsymbol{v}(\boldsymbol{M}_i) = \frac{\sum_{\boldsymbol{i}=1}^{\boldsymbol{n}} (\boldsymbol{n}(\boldsymbol{Mi}) - \boldsymbol{m}(\boldsymbol{M_i}))^2}{\boldsymbol{N}} \tag{5}$$

where $v(M_i)$ is the i^{th} morpheme variance, $n(M_i)$ is normalized morpheme frequency in a document, $m(M_i)$ is mean value, and N is the total number of distinct morphemes in the document.

$$m(M_i) = \frac{\sum_{i=1}^{n} MF}{N} \tag{6}$$

where MF is morpheme frequency and N is the number of words in a document.

Entropy: This is used to measure the information value e of each morpheme in the corpus. This method is based on the amount of information a morpheme carries. Stopwords are known to have low explanatory values [3]. If the entropy value of a word is high, then the information value of the word is low. The entropy value of each morpheme in the corpus is calculated as:

$$e(M_i) = \sum_{i=1}^{n} p(M_i).\log\frac{1}{p(M_i)} \tag{7}$$

where $p(M_i)$ is the probability of morpheme frequency and is calculated by dividing the morpheme frequency via the total number of morphemes in the document.

Extraction of Stopwords: The frequency, mean, variance and entropy values of each morpheme in the corpus are compared against other morphemes in the corpus to select stopwords. The intersection of the top n words from these statistical information are selected as stopwords. Finally, the stopword list contains stem-based and root-based stopwords. The top 250 morphemes from the four stopwords lists (Table 2, Table 3, Table 4 and Table 5) are merged together to generate the aggregated corpus based stopword list. These morphemes make up a large fraction (more than 60%) of the Amharic text documents. However, the amount of information carried by these morphemes is negligible. The final Amharic stopword list is generated from the intersection of the frequency, mean, variance and entropy stopword lists with few linguistics experts' analyzed words. The list contains 222 words. Variants of a stopwords are not included in the constructed stopwords list. Their length is from one to five characters. They occur much more frequently than other morphemes. They include prepositions (e.g. ወደ/*wədə* 'to'/, እንደ */ʔinidə* 'such as'/, ስለ /*silə* 'about'/, እስከ /*ʔisikə* 'up to'/, በ /*bə* 'by'/, ከ /*kə* 'from'/, etc.), conjunctions (e.g. እና /*ʔina* 'and'/, ወይም /*wəyim*\xE2 \x80\x98or'/, ይሁን እንጂ /*yihun ʔinidʒ* 'however'/, ምክንያት /*miknijat* 'because'/, እዚህ /*ʔizih* 'this'/, ይልቅ /*jilik* 'instead'/, etc.), negative markers አል /*ʔəl* 'not'/, ም /*mə*/), indefinite articles (አንድ /*ʔənid* 'an'/), auxiliary verbs ኦ-ል /*ʔ-l* 'say'/,ን-በ-ር /*n-b-r* 'was'/, etc.), ወዘተ /*wəzətə* 'and so on'/, etc. These and the like words are found at the top of the four lists (frequency, mean, variance, and entropy).

4 Result and Discussion

4.1 Result

The overall distributions of morphemes in the corpus are measured using frequency, mean, variance, and entropy values as presented as follows. From a statistical point of

view, the intersection of top n morphemes across various measurements are considered as stopwords. Accordingly, a total of 222 stopwords are identified and included in our list of stopwords. The list of sample morphemes with highest morpheme and collection frequencies are shown in Table 2. The distribution of stopwords across many documents in the corpus is presented in Table 2. Their document and collection frequency are higher than non-stopwords. Stopwords are found in majority of the documents with high collection frequency. Morphemes that had low frequencies are excluded from stopwords.

Table 2. Top 10 morphemes with highest document and collection frequency

Morpheme	Document frequency	Collection frequency
የ /*jə* 'of'/	5,737	190,726
በ /*bə* 'by'/	5,733	139,870
ኡ /*ʔu* 'they'/	5,731	141,646
ኧ /*ʔə* 'he'/	5,727	105,751
ው /*wɨ* 'the'/	5,717	99,033
አል /*ʔəl* 'not'/	5,715	72,269
አ /*ʔə* 'he'/	5,709	73,261
መ /*mə* nominator/	5,708	91,533
ተ /*tə* passivizer/	5,707	84,031
ን /*nɨ* accusative/	5,703	92,342

Amharic stopwords have high mean probability values compared to the majority of non-stopwords. Samples of stopwords with the highest mean probability are shown in Table 3.

Table 3. Top 20 morphemes with highest mean values

Morpheme	Mean
የ /*jə* 'of'/	0.05810
በ /*bə* 'by'/	0.04196
ኡ /*ʔu* 'they'/	0.04079
ኧ /*ʔə* 3pms/	0.02915
ው /*wɨ* 'the'/	0.02836
መ /*mə* nominator/	0.02570
ን /*nɨ* accusative/	0.02554
ተ /*tə* passivizer/	0.02451
አል /*ʔəl* 'not'/	0.01565
ኦች /*ʔotʃ* 'many'/	0.01549

Morpheme	Mean
አት /*ʔəti* 'not'/	0.01434
አስ /*ʔəs* 'be the cause of'/	0.01011
ሀ-ን /*h-n* 'happen'/	0.00888
አቸው /*ʔətʃəw* 'are'/	0.00806
ድ-ር-ግ /*d-r-g* 'act'/	0.00738
ያ /*ja* 'that'/	0.00712
እንደ /*ʔɨndə* 'like'/	0.00675
ላይ /*laj* 'on'/	0.00553
እንዲ /*ʔɨndi* 'as'/	0.00329
ነው /*nəw* 'is'/	0.00326

Stopwords have highest variance probability value than almost all content bearing morphemes. They are located at the top of the ranked list. Table 4 shows top 10 morphemes with highest variance values.

Table 4. Top 10 morphemes with highest variance values

Morpheme	Variance	Morpheme	Variance
የ /*jə* 'of/	0.00132	ው /*wɨ* 'the' /	0.00038
በ /*bə* 'by'/	0.00072	ን /*nɨ* 'we'/	0.00031
ኡ /*ʔu* 'the'/	0.00069	ም /*mə* Inf/	0.00031
አ /*ʔə* 3psm/	0.00055	ተ /*tə* pas/	0.00029
ኧ /*ʔə* 3psm/	0.00045	ኦች /*ʔotʃɨ* 'many'/	0.00018

Stopwords have highest entropy value than non-stopwords. The top *n* morphemes with the highest entropy value are extracted as candidate for stopwords. Table 5 shows top 10 morphemes with highest entropy values.

Table 5. Top 10 morphemes with highest entropy values

Morpheme	Entropy	Morpheme	Entropy
የ /*jə* 'of/	38.01462	ተ /*tə* pas/	20.68732
በ /*bə* 'by'/	31.38035	አ /*ʔə* 3psm/	20.17104
ኡ /*ʔu* 'the'/	30.22907	ም /*mə* Inf/	18.96849
ው /*wɨ* 'the'/	24.00580	አል /*ʔəl* 'not'/	16.54033
ን /*nɨ* 'we'/	22.48845	ለ /*lə* 'to'/	14.58708

4.2 Discussion

Based on their nature they can be classified into three. The first types of Amharic stopwords exist by themselves and can accept prefixes and suffixes. For instance, the stopword ሌላ /*lela* 'other'/ can take the prefix የ /*jə* 'of'/ and the suffix ኦች /*ʔəotʃɨ* 'many'/ become የሌሎች /*jəlelotʃ* 'any others'/. The second types of stopwords exist as standalone words but do not take affixes, e.g. ወዘተ /*wəzətə* 'and so on'/, ወይም /*wəyim* 'or'/, etc. The third types of stopwords exist as part of Amharic words and act as prefix or suffix. For instance, the words ከጎንደር /*kəgondər* 'from Gondar'/ and በመኪናው/*bəməkinaw* 'the car'/ contain the prepositions በ /*bə* 'by'/ and the suffix ው /*wɨ* 'the'/, respectively. As the meaning of stopwords indicated in Tables 2, 3, 4 and 5, they are similar to English language semantically and functionally except linguistic differences. Unlike morphologically simple languages, identification of Amharic stopwords is not an easy task. The challenges are

the existence of many forms for a stopword. The majorities of stopwords undergo complex morphological process, and merge with each other or other words to form new words as shown in Table 6.

Table 6. Examples of morphological changes in stopwords

Word	Morpheme	Morphological information
በውስጥና	በ_ውስጥ_ና /*bə_wist'_ʔi*na/	preposition-stem-conjunction
የእነዚህ	የ_እነ_እዚህ /*jə_ʔinə_əzih*/	genitive-preposition- demonstrative
በመሆኑ	በ_መ_ህ-ን_ኡ /*bə_mə_h-n_ʔu*/	preposition-nominalizer-root-definite
ከሰውና	ከ_ሰው_ና /*kə_səw_ʔi*na/	preposition-stem-conjunction

As shown in Table 6, many of Amharic stopwords affix with other stopwords or non-stop words. They might change their forms depending on their context. Therefore, it is not possible to find and remove all Amharic stopwords directly before morphological analysis. Hence, to see their distribution and information content in the corpus, their frequency (document and collection), mean, variance and entropy need to be computed after morphological analysis. Previously created stopword lists contain simply small number of variants of stopwords. However, it is very difficult, if not impossible, to include all variants into a list of stopwords. The following examples show how various word forms can be formed from a single stopword.

ውስጥ፡ ውስጥና, ውስጥም, የውስጥ, ለውስጥ, በውስጥ, ከውስጥ, የውስጥና, ከውስጥም, ከውስጥና, ለውስጥና, በውስጥም, ውስጥን, በውስጥህ, በውስጥና, በውስጥሽ, የውስጥም, ውስጥህ, ውስጥስ, ከውስጣቸው, በውስጡ, ከውስጧ, etc.

መካከል፡ መካከልም, በመካከል, በመካከላቸው, ከመካከላቸው, በመካከሏ, መካከልና, በመካከልሽም, ከመካከል, በመካከልም, የመካከልኛው, በመካከሉ, etc.

በኩል፡ በበኩላቸው, የበኩላቸውን, የበኩሉን, በበኩሉ, በኩልም, የበኩሏን, በበኩሏ, በየበኩላቸው, የበኩላችንን, በበኩላችሁ, የበኩላቸው, የበኩሌን, በበኩላቸው, በበኩሉ, የበኩላችሁን, በኩልና ,etc.

ላይ፡ በላይ, ላይም, ከላይ, ላይና, በላይም, ባላይ, የላይኛው, ወደላይ, በላይና, የላይኛውን, በላይዋ, በላይኛው, የላይ, ላይኛው, በበላይ, እላይ, ላይኛውን, ከበላይ, የላይኛውና , ለበላይ, etc.

ብቻ፡ ብቻውን, ብቻቸውን , ለብቻ, ለብቻው, ብቻም, ብቻየን, ለብቻቸው, ለየብቻ, ብቻዋን, ለብቻየ, ብቻህን, ብቻችንን, ብቻውንም, ብቻችሁን, ለብቻዋ, ለብቻችን, etc.

ኋላ፡ በኋላ, በኋላም, ወደኋላ, ከኋላ, የኋላ, ከኋላው, በኋላና, ኋላም, ከኋላቸው, ከኋላየ, ኋላው , በኋላማ, በኋላው, ከኋላህ, ከኋላዋ, ከኋላችን, በኋላስ, ኋላየ, የኋላው, ኋላዋ, ኋላውም, etc.

ሌላ፡ ሌሎች, በሌሎች , በሌላ , ከሌሎች, ሌላው, ሌላኛው, ለሌሎች, ሌሎችም, በሌሎችም, የሌላቸው, የሌሎች, ሌሎችንም, ከሌላው, ሌሎችን, ለሌሎችም, ከሌሎችም, የሌሎችን, እንደሌሎች, ለሌላ, ሌላውን, ለሌላው, በሌላው, በሌላም, ከሌላ, የሌላቸውን, ሌላም, የሌሎችንም, ወደሌሎች, የሌሎችም, በሌሎችም, ስለሌሎች, ከሌሎች, ሌላዋ, የሌላ, የሌላውን, ለሌላቸው, በሌላቸው, በሌላኛው, etc.

እዚህ፡ በዚህ, ከዚህ , በዚህም, እነዚህ, ለዚህ, ለዚህም, ስለዚህ, እነዚህን, ከዚህም, እዚህ, የዚህ, በእነዚህ, ከእነዚህ, ከነዚህ, እንደዚህ, ስለዚህም, ከእነዚህም, ወደዚህ, በነዚህ, የእነዚህ, ለእነዚህ, የዚህን , በዚህች, የእነዚህን, etc.

In our work, such morphological variants of a stopword are represented by one common form that is stored in the stopword list. Some Amharic stopwords have both

stem and root forms. However, only root-based approach can conflate all variants of a stopword to one common form. For example, the stopword ንብር /*n-b-r*/ has two stem forms which are ነበር- /*nəbər*- 'was'/ and ነባር- /*nəbar*- 'was'/. Thus, we use root forms to represent Amharic stopwords. Analysis of the Amharic stopwords reveals that most of them are affixes without particular semantic information. They are rather used for syntactic purpose such as definite articles, prepositions, conjunctions, negative markers, etc. appearing mostly as prefixes and suffixes.

5 Evaluation of Stopwords

The constructed generic Amharic stopword list contains morphemes and may be used in various fields. The stopwords are evaluated based on their effect on term weighting and retrieval effectiveness. Experiments were conducted to test the impact of stopwords on term weighting and retrieval effectiveness as follows.

The Effect of Stopwords on Term Weighting: The effect of Amharic stopwords on term weighting is investigated and tested. Term weighting is used in IR field to extract the most relevant terms of documents. The Term Frequency-Inverse Document Frequency (TFIDF) of every word in each document was computed to evaluate the importance of each word to represent document content. The term weighting of stopwords and some non-stop words were evaluated using Amharic IR test collection [25]. The statistical analysis and the comparison between the existence and removal of stopwords on term weighting are presented in Table 6 as follows. Documents and terms are selected randomly in the corpus. It can be seen that the term weighting of non-stop words slightly increases after elimination of stopwords. This means that the importance of terms to represent the subject matter in each document increases. The TFIDF values of stopwords are zero or close to zero. For example, stopwords such as የ /*jə*/, በ /*bə*/, ኡ /*ʔu*/, ኧ /*ə*/, ው /*wi*/, አል /*ʔəl*/, አ /*ʔə*/, መ /*mə*/, ተ /*tə*/ and ን /*ni*/ have zero TFIDF values whereas stopwords such as እና /*ʔina*/, ኦች /*ʔotʃ*/, አት /*ʔət*/, አስ /*ʔəs*/ , እንደ/*ʔinidə*/, አቸው /*ʔətʃəw*/, እንዲ /*ʔinidi*/, ዎች /*wotʃ*/, ነው /*nəw*/ and ላይ /*laj*/ have values ranging from 0.01 to 0.07. This means that they are not significant to describe the content of documents.

The Impacts of Stopwords on Retrieval Effectiveness: The effect of stopwords removal on Amharic IR retrieval effectiveness is tested on Amharic IR test collection [25]. The evaluation was done using Lemur and trec_eval tools on language modeling approach. Figure 2 shows retrieval effectiveness with and without stopwords on stem-based and root-based approaches. The top two graphs (labeled in red and blue) represent root-based and stem-based retrieval without stopwords while the bottom two graphs (green and yellow) represent root-based and stem-based retrieval with stopwords. The evaluation is made using the same Amharic IR test collection [25]. It can be seen that there is a significant difference between Amharic retrieval with and without stopwords on stem-based and root-based approaches. The bottom graph (in yellow) and the second top graph (in blue) indicate retrieval with and without stopwords on stem-based approach, respectively. The third top graph (in green) and top graph (in red) are retrieval with and without stopwords on root-based approach, respectively. Root-based retrieval

is better than stem-based retrieval. This is because root-based approach is best to conflate all variants but not stem-based approach.

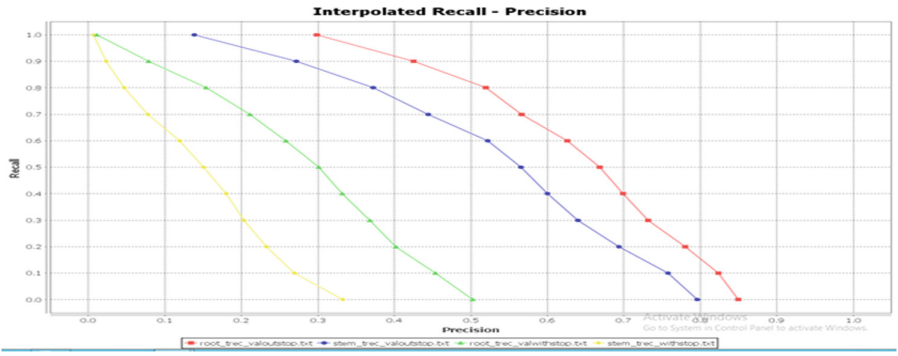


Fig. 2. Retrieval effectiveness with and without stopwords

In addition to retrieval effectiveness, the impact of stopwords on index size is evaluated. The number of morphemes decreases significantly after stopwords removal. The numbers of morphemes before and after stopwords removal are 3,399,172 and 1,316,504, respectively. The effect of stopwords on the size of index file is shown in Table 7. Removal of stopwords has reduced the size of index file, which minimizes the time of processing index file.

Table 7. Index size with and without stopwords on root-based corpus

Index type	Corpus size in MB	Index size
With stopwords	33.0 MB	25.1 MB
Without stopwords	24.1 MB	15.1 MB

The availability of standard stopword is a major factor in developing different Amharic applications. We believe that a resource developed for a research purpose should be easily available to researchers and developers. Hence, the developed stopword list in this research is made publicly accessible online for future researches. So it enables researchers and developers to build their systems at minimal cost.

6 Conclusion

Amharic is one of the under-resourced languages facing lack of NLP resources, tools and corpora. We present an Amharic stopword list created by considering the semantics and characteristics of stopwords in the language, and by analyzing their distribution in a corpus. The applicability of stopwords is systematically evaluated using Amharic

IR system. Thus, the stopword list is believed to be a generic resource for other NLP applications as well. The stopword list is made publicly available for the research community and can be accessed through a request made to the corresponding author at: tilahun.yeshambel@uog.edu.et.

References

1. Gerlach, M., Shi, H., Amaral, L.A.N.: A universal information theoretic approach to the identification of stopwords. Nat. Mach. Intell. (2019)
2. Antoine, B.: Understanding and customizing stopword lists for enhanced patent mapping, vol. 29, no. 4, pp. 308. Elsevier (2007)
3. Feng, Z., Lee, W., Xiaotie, D., Song, H., Sheng, W.: Automatic construction of Chinese stopword list. In: Proceedings of the 5th WSEAS International Conference on Applied Computer Science, pp. 1010–1015, Hangzhou (2006)
4. Alemu, A., Lars, A.: Amharic-English information retrieval. In: Workshop of the Cross-Language Evaluation Forum for European Languages, pp. 43–50. Springer (2006).
5. Khalifa, C., Rayner, A.: An Automatic construction of Malay stopwords based on aggregation method, Singapore, pp. 180–189. Springer Nature Pte Ltd. (2016)
6. Ibrahim, A.: Effects of stopwords elimination for Arabic information retrieval: a comparative paper. Int. J. Comput. Inf. Sci. **4**(3), 119–133 (2006)
7. Ounis, I., Amati, G., Plachouras, V., Macdonald, C., Johnson, D.: Terrier: a high performance and scalable information retrieval platform. In: SIGIR Open Source Workshop 2006 Seattle, Washington (2005)
8. Rani, R., Lobiyal, D.: Automatic construction of generic stopwords list for Hindi text. Proc. Comput. Sci. **132**(Iccids), 362–370 (2018)
9. Sadeghi, M., Vegas, J.: Automatic identification of light stopwords for Persian information retrieval systems. J. Inf. Sci. **40**(4), 476–48 (2014)
10. Yeshambel, T., Mothe, J., Assabie, Y: Evaluation of corpora, resources and tools for Amharic information retrieval. In: ICAST2020. springer, Bahir Dar (2020a)
11. Mindaye, T., Redwan, H., Atnafu, S.: Searching the Web for Amharic content. Int. J. Multimed. Process. Technol. (JMPT) **1**(1), 318–325 (2010)
12. Samuel, E., Bjorn, G.: Classifying Amharic news text using self-organizing maps, vol. 71 (2005)
13. Alemayehu, N., Willett, P.: Stemming of Amharic words for information retrieval. Literary Linguist. Comput. **17**(1), 1–17 (2002)
14. Alemayehu, N., Willett, P.: The effectiveness of stemming for information retrieval in Amharic. Prog.: Electron. Libr. Inf. Syst. **37**(4), 254–259 (2003)
15. Asefa, G.: Ontology-based semantic indexing for Amharic text in football domain. Master's thesis, Addis Ababa University, College of Natural Science, Ethiopia (2013)
16. Hirpa, A.: Probabilistic information retrieval system for Amharic language. Master's thesis, Addis Ababa University, School Of Information Science, Ethiopia (2012)
17. Mengistu, B.: N-gram-based automatic indexing for Amharic text. Master's thesis, Addis Ababa University, School Of Information Science, Ethiopia (2002)
18. Mengiste, B.: Automatic ontology learning from unstructured Amharic text. Master's Thesis, Addis Ababa University, College Of Natural Sciences, Department Of Computer Science, Ethiopia (2013)
19. Asker, L., Argaw, A., Gambäck, B.: Applying machine learning to Amharic text classification. WOCAL 5: 5th World Congress of African Linguistics, pp. 7–11. Addis Ababa University, Ethiopia (2006)

20. Assabie, Y.: Development of Amharic morphological analyzer. Technical report, Ministry of Communication and Information Technology, Addis Ababa (2017)
21. Yifru, M., Wolfgang, M.: Morphology-based language modeling for Amharic. Ph.D.'s thesis, University of Hamburg, departments of informatics, German (2010)
22. Mulugeta, W., Michael, G.: Learning morphological rules for Amharic verbs using inductive logic programming. Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL8/AfLaT2012), pp. 7–12 (2012)
23. Yimam, B.: Yamarigna Sewasiw (Amharic Grammar). 2nd edn. CASE, Addis Ababa (2001).
24. Stefano, F., Floriana, E., Domenico, G.: Automatic learning of linguistic resources for stopword removal and stemming from text. Elsevier Proc. Comput. Sci. **38**, 116–123 (2014)
25. Yeshambel, T., Mothe, J., Assabie, Y.: 2AIRTC: the amharic adhoc information retrieval test collection. In: Arampatzis, A., et al. (eds.) CLEF 2020. LNCS, vol. 12260, pp. 55–66. Springer, Cham (2020b). https://doi.org/10.1007/978-3-030-58219-7_5