Amharic Document Representation for Adhoc Retrieval

Tilahun Yeshambel, Josiane Mothe, Yaregal Assabie

Amharic Document Representation for Adhoc Retrieval

Tilahun Yeshambel[1], Josiane Mothe[2] and Yaregal Assabie[3]
[1]*IT PhD program, Addis Ababa University, Addis Ababa, Ethiopia*
[2]*INSPE, Univ.de Toulouse, IRIT, UMR5505 CNRS, Toulouse, France*
[3]*Department of Computer Science, Addis Ababa University, Addis Ababa, Ethiopia*
tilahun.yeshambel@uog.edu.et, josiane.mothe@irit.fr,yaregal.assabie@aau.edu.et

Keywords: Adhoc Retrieval, Amharic, Complex Morphology, Stem, Root.

Abstract: Amharic is the official language of the government of Ethiopia currently having an estimated population of over 110 million. Like other Semitic languages, Amharic is characterized by complex morphology where thousands of words are generated from a single root form through inflection and derivation. This has made the development of tools for Amharic natural language processing a non-trivial task. Amharic adhoc retrieval faces difficulties due to the complex morphological structure of the language. In this paper, the impact of morphological features on the representation of Amharic documents and queries for adhoc retrieval is investigated. We analyze the effects of stem-based and root-based approaches on Amharic adhoc retrieval effectiveness. Various experiments are conducted on TREC-like Amharic information retrieval test collection using standard evaluation framework and measures. The findings show that a root-based approach outperforms the conventional stem-based approachthat prevails in many other languages.

1 INTRODUCTION

Searching digital information on the Web or document collection has long become part of the human daily life. Information Retrieval (IR) is the task of searching relevant documents to a user query from document collection. Both the research community and the industry have been very active in this field for more than 60 years (Sanderson and Croft, 2012). Nowadays, IR has gained much attention due to the explosion of digital data and the need of accessing relevant information from huge corpus quickly and accurately.

IR systems work based on documents representing natural languages, and consequently, the characteristics of a given language affects the whole process of IR (Moukdad, 2002). Thus, natural language processing (NLP) has attracted the attention of IR communitysince a long time (Smeaton, 1992; Jackson and Moulinier, 2007; Cambria and White, 2014). For example, NLP applications and resources provide a means to find better representative terms for indexing and query terms that improve search results. This calls for the need of dealing with language specific issues to improve the performance of IR systems. The morphology, orthography, tokenization, syntax, semantics, and compound splitting of a language are some of the issues to be considered while developing IR systems. It has long been understood that linguistic variation has significant impact on IR effectiveness as it leads to the omission of relevant documents to users' queries(Moukdad, 2002). Many languages have different forms generated from a single word due to morphology and orthography. Identifying the basic units of words is more difficult for morphologically complex languages than for simple languages. Performing simple matching between words generated from the same root is not applicable to capture similarity.

Thus, in order to come up with an effective IR system, one has to deal with the complex characteristics of the language. One of the key features of Amharic is its complex morphology which itself leads to complex grammatical structure. As a result, finding better representations for documents and queries has been an issue of theoretical discussion in Amharic IR. The forms that can be considered for document and query representation are stems and roots.In many languages, the use of surface forms of words to represent documents and queries is not taken into consideration due to the proliferation of words that can be generated from a single root form. This issue

is imperative in the development Amharic IR. Although some efforts have been made to develop Amharic IR systems using stems, their effectiveness with respect to the use of various forms has not been systematically analyzed thus far. Therefore, thisresearch analyzes the use of stems and roots for content representation and investigates their effects on Amharic IR.

The rest of this paper is organized as follows. Section 2 describes Amharic language and its morphology. Section 3 discusses related work and Section 4 presents how documents and queries are represented in Amharic IR system. Experimental results and evaluation are discussed in Section 5. In Section 6, we make conclusion along with the way forward in Amharic IR.

2 AMHARIC LANGUAGE

Amharic is the official language of the government of Ethiopia. Although several languages are spoken in Ethiopia, Amharic is spoken as a mother tongue by a sizeable proportion of the country's population currently estimated to be over 110 million. Among the Semitic language family, it is the second most spoken language in the world, next to Arabic. Due to its historical significance and official status, Amharic has been serving as the *lingua franca* of the country since a long time. As a result, many literary works, government documents, educational materials, religious literary works, etc. are predominantly produced in Amharic. Amharic uses Ethiopic script for writing having 34 base characters (with a vowel ኧ /*ə*/), each of which are modified to have six other orders representing vowels in the order of ኡ /*u*/, ኢ /*i*/, ኣ /*a*/, ኤ /*e*/, እ /*ɨ*/, and ኦ /*o*/.

Like other Semitic languages, complex morphological processes are carried out on Amharic word classes such as verbs, nouns and adjectives (Yimam, 2001). Amharic verbs are the most complex word classes and can be generatedby attaching affixeson verbal stems. On the other hand, verbal stems can be generatedfrom verbal roots by inserting vowels between radicals. For example, the verbal stem ገደል-/*gədəl-*/ is derived from the verbal root ግ-ድ-ል/*g-d-l*/. Moreover, verbal stems (e.g. ተገደል-/*təgədəl-*/) can be derived from other verbal stems (e.g. ገደል-/*gədəl-*/) by affixing morphemes. The verb formation process is usually completed by attaching a verbal stem with person, gender, number, case, tense/aspect and mood markers. For example, from the verbal stem ገደል- /*gədəl-*/ the following verbs can be generated: ገደልኩ/*gədəlku*'I killed'/, ገደልኩህ/*gədəlkuh*'I killed you'/, ገደልን/*gədəln*'we kill'/, ተገደልኩ/*təgədəlku*'I was killed'/, ገደለች/*gədələtʃ*'she killed'/, etc. As verbs are marked for subject and object, they alone can representa completesentence.For example, the word አልሰበረንም/*ʔəlsəbərənɨm*'he did not break us'/, which is constructed from the morphemes *ʔəl-səbər-ə-nɨ-m*, is a complete sentence with the following linguistic information:*ʔəl-…-m* /not/, *-səbər-* /did break/, *-ə-* /he/ and *-nɨ-* /us/. Accordingly, thousands of verbs can be derived from a verbal root through a complex morphological process carried out by attaching a combination of person, case, gender, number, tense, aspect, mood and others (Abate and Assabie,2014;Assabie, 2017).

Based on a morphological structure, Amharic nouns and adjectives can be either derived or non-derived. For example, the word መሬት/*məret*'earth'/ and ዛፍ/*zaf*'tree'/ are non-derived nouns whereas words like ስብራት /*sɨbɨrat*'the state of being broken'/ and ደግነት/*dəgɨnət* 'generosity'/ are nouns derived from the verbal root ስ-ብ-ር/*s-b-r* 'to break'/ and the adjective ደግ/*dəg* 'generous'/, respectively. Derived nouns are generated from other word classes though morphological processes. In general, Amharic nouns can be derived from verbal roots, adjectives and other nouns by affixing vowels or bound morphemes. Derived adjectives can be formed from verbal roots by infixing vowels between consonants (e.g. ክ-ብ-ድ/*k-b-d* 'to become heavy'/ → ከባድ/*kəbad* 'heavy'/), nouns by suffixing bound morphemes such as -ኧኛ /*ʔəɲa*/ (e.g. ጉልበት/*gulbət* 'power'/ → ጉልበተኛ/*gulbətəɲa* 'powerful') and verbal stems by prefixing or suffixing bound morphemes (e.g. ደካም-/*dəkam-*/ → ደካማ/*dəkama* 'weak'/). Although the morphological process of derivation of nouns and adjectives is complex by itself, even more complexity arises from their inflections. Amharic nouns and adjectives are inflected for number by suffixing -ኦች/*-ʔotʃ*/ or -ዎች/*-wotʃ*/, definiteness by suffixing -ኡ/*-ʔu*/ or -ዉ/*-wu*/, objective case by suffixing -ን/*-n*/, possessive case by suffixing different morphemes depending on the subject, and gender by suffixing -ኢት/*-ʔit*/. These inflections can appear alone or in combination at the same time, along with prepositions and negation markers which lead to the generation of thousands of word forms from a single noun or adjective. For example, ያለባለቤቶቹ/*jaləbaləbetotʃu* 'without the owners of the house'/ is generated from the morphemes *jə-ʔələ-balə-bet-otʃ-u*(*jə*preposition 'of/with', *ʔəl*negation marker'not/without', *balə*possessive marker'owner of', *bet*noun 'house', *otʃ* plural marker, and *u*definite

marker 'the') where the core morpheme is the noun ቤት/*bet* 'house'/.

3 RELATED WORK

Semitic languages are known to pose unique challenges in the development of NLP applications due to their complex morphologies. These challenges are propagated to the development of IR systems since the effectiveness of IR systems depends on the availability of various NLP tools and resources. In this section, we discuss the techniques and NLP resources used to develop IR systems for Semitic languages in general.

Arabic is the largest of the Semitic language family. Arabic IR systems have a long history (Ambati*et al*., 2008;Larkey *et al*., 2007;Darwish and Magdy, 2014). For example, Al-Hadid*et al*. (2014) developed a neural network-based model where documents and queries are represented using stems and their similarity is computed using cosine similarity. Musaid (2000) investigated the effectiveness of word-based, stem-based, and root-based representation of documents and queries.The word-based and stem-based approaches miss relevant documents while root-based approach retrieves non-relevant documents. Moukdad (2002) conducted a resarch to compare the effects of stem and root on Arabic IR. The retrieval effectivness of stem and root were evaluated on search engine. The results of the expermints indicas stemming is more effective than root.Larkey *et al*.(2007) investigated the effects of light stemming (removal of prefix and suffix) on Arabic IR. A comparasion between stem-based and root-based retrieval was performed. The finding indicates light steming outperforms than root and other stemmers which are based on detail morphological analysis.Abdusalam (2008) presented an Arabic text retrieval technique using lexicon-based light stemming. The studyevaluated the effectiveness of lexicon-based light stemming, Arabic patterns, root, expanding query and filtering foreign words using n-grams. According to the results, the preprocessing techniques like normalization, stopword removal and light-stemming improve retrieval results whereas n-grams and roots decrease the performance. The lexicon-based stemming and the relevance feedback approaches perform better than light-stemming approach alone. Ali *et al*. (2020)investigated the effect of morphological analysis on Arabic IR. A rule-based stemmer was used to extract the root/stem of words to be used as indexing and searching terms. The resultsshowed slight improvement on IR effectiveness due to the stemmer.

Hebrew is one of the Semitic languages spoken mainly in Israel. Carmel and Maarek(1999) presented a morphological disambiguator based on a statistical approach that takes advantage of an existing morphological analyzer. The approach is context-free and was used for query analysis and linguistic indexing of text documents. Instead of words, the morphological patterns were used for disambiguation. The statistical morphological disambiguator returns only the best base form(s), or lemma(s). It makes the decisions of the most likely set of analyses based on the frequency of the morphological patterns associated with the analyses of the input word. The disambiguator was tested by integrating with the Hebrew search engine. It conflates all inflectional forms andthe performance of the search engine increased. Ornan(2002) designed Hebrew search engine by applying a rule-basedmorphological analysis. The design of the search engine takes into account the construction of a morphological, syntactic and semantics analyzer. The search engine eliminates words unsuited both to the syntax and the semantic of a sentence.

Although Amharic is significantly used in Ethiopia, the status of IR system development for the language is relatively at rudimentary level. Alemayehu and Willett (2003) studied the retrieval effectiveness of word-based, stem-based, and root-based approaches on Amharic language. The experiments were carried out by running 40 queries on 548 documents using OKAPI systemand the study concludes that stem-based retrieval is slightly better than root-based. Similarly, Mindaye*et al*. (2010) developed an Amharic search engine using stems. The system was tested with11 queries on 75 news documents. The average precision and recall values were 0.65 and 0.95, respectively using OR operatorin between query terms, and 0.99 and 0.52, respectively for AND operator.Argaw*et al*. (2004) developed dictionary-based Amharic-English IR system. Documents and queries were represented using Bag-Of-Words (BOW). Stopwords were removed using Inverse Document Frequency (IDF) and stopwordslist. The average precisions of 0.3615 and 0.4009 were achieved using IDF and stopword list, respectively.Argaw *et al*. (2006) build dictionary-based Amharic-French IR system with and without word sense discrimination using BOW approach.Stemming was applied to remove prefix and suffix. The experiments were conducted on SICS and Lucene search engines. Stopwords were removed by using IDF. The result of SICS is better

than Lucene. The word sense descrimination performs slightly better than non discrimination.

While there are several studies that focused on the development of IR systems for Semitic languages, most of them have followed the techniques employed for morphologically simple languages like English. This has not produced the desired retrieval result as documentscould not be represented appropriately. Only fewstudies have tried to consider the issue of document representation in a systematic way. In our approach, we address this crucial issue of document representation in the development of Amharic IR.

4 DESIGN OF AMHARIC IR

The main objective of this work is to systematically identify the optimal representations for documents (and queries) in Amharic IR. It focuses on the selection of the structures of terms and stopwords based on the morphological characteristics of the language. Taking these issues intoaccount, we also propose an Amharic retrieval system which is slightly different from the basic architecture of IR systems. In our case, stopwords are removed after the application of morphological analysis on documents and queries as shown in Figure 1.Both documents and queries pass through the same preprocessingtasks that involve language specific tokenization,character normalization, and removalof punctuationmarks. Text preprocessing is followed by morphological analysis which is performed on documents and queries to select appropriate terms for document representation. Morphological analysis is among the key tasks in our IR system as it helps to remove stopwords from documents and queries.

4.1 Preprocessing

Preprocessing includes tags removal, tokenization, character normalization and punctuation mark removal.Tokenization is done using space andpunctuationmarks as delimiters of words.Character normalizationis made to represent various characters having similar pronunciation using a single grapheme.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/}.Furthermore, the fourth orders characters {ሃ /ha/, ሓ /ha/, ኃ /ha/ and ኻ /ha/} and {ኣ/ʔa/ and ዓ/ʔa/} have similarity phonemes with the correspondence base character.Therefore, Amharic character normalizationinvolves mapping of ሀ /ha/,ሃ /ha/, ሐ /ha/, ሓ /ha/, ኃ /ha/,and ኃ /ha/toሀ/ha/; አ /ʔa/, ኣ/ʔa/, ዐ /ʔa/,and ዓ /ʔa/ to ኣ/ʔa/, etc. For example, the word ስለጸሀፊዋ can also be written as ሥለጸሀፊዋ, ስለጸሓፊዋ, ሥለጸሓፊዋ, ሥለጸኀፊዋ, ስለጸኀፊዋ, ሥለጸሓፊዋ,ስለፀሀፊዋ,ስለፀሓፊዋ,ሥለፀኃፊዋ,ስለፀኻፊዋ,etc. although some of them rarely appear in a text.

4.2 Morphological Analysis

Documents and user information need should be represented appropriately using terms that will be used later for matching query with document. It is to be noted that indexing terms are weighted based on theword frequency. In IR, most often, the variants of a word are conflated during indexing into a single form.It has the advantage of making the calculation of indexing term frequency straightforward.

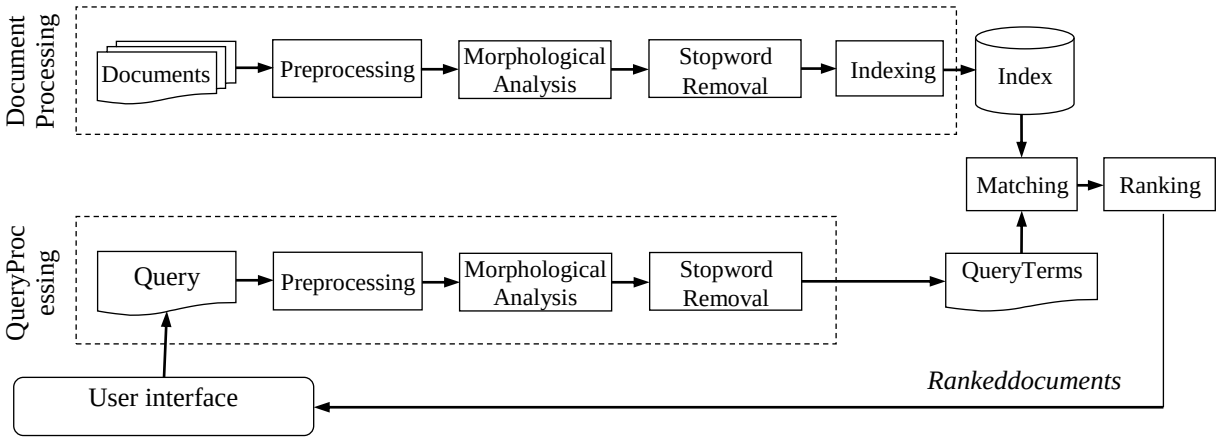

Figure 1: Basic architecture of the proposed Amharic IR system.

Therefore, in this research,we study the feasibility of stem-based and root-baseddocument representation with respect to their effectiveness for Amharic IR. Since well-designed Amharic morphological analyzer is not available, we design semi-automatic annotationto morphologically segment words in documents.

4.2.1 Stem-Based Morphological Analysis

One of the bases of Amharic words are stems. A large number of words are formed by attaching affixes to stems. Therefore, morphological analysis should be carried out to extract the stem from the rest of morphemes. For example, the morphological structures of the primary noun በመንገዶቻችን/*bəməngədocatʃin*'by our roads'/and the adjective የደጋጎች/*jədəgagotʃ*'of generous'/are shown as follows.

በመንገዶቻችን	የደጋጎች
በ_መንገድ_አች_አችን	የ_ደግ_አች
pre-stem-pl-1,pl[1]	*gen-stem-pl*
by-road-many-our	of-generous-many

Similarly, variants of adjectives and nouns derived from primary nouns are mapped into their common stems of nouns. For example, the derived noun ስለልጅነቴና/*siləlidʒinətena*'and about my childhood'/ and the derived adjective ዓለማዊ/*ʔələmawi*'worldly'/ are morphologically segmented as follows.

ስለልጅነቴና	ዓለማዊ
ስለ_ልጅ_ነት_ኤ_ና	ዓለም_አዊ
pre-stem-nom-1,s-con	*sem-adj*
about-child-being-my-and	world-suf

Amharic verbs undergo complex morphological process. Verbs are marked for person, gender, number, tense, subject, object, and negation by attaching a series of affixes. For example, the word ፈለገቻችሁ/*fələgətʃatʃihu*'she wanted you'/and the verb ከአልተስማማናቸውም/*kəʔəltəsmamanatʃəwim*'and if we are not comfortable for them'/ is analysed as follows.

ፈለገቻችሁ	ከአልተስማማናቸውም
ፈለግ_ሸች_አችሁ	ከ_አል _ስማም_ን_አቸው_ም
stem-sub-obj	*pre-neg-stem-1,pl-3,pl-neg*
search-she-you	from-not-comfort-we-they-not

Amharic has both basic and derived (causative, passive, infinitive and reduplicative) types of verbal stems.Causative stems are formed using the prefixes አ-/*ʔə-*/ and አስ-/*ʔəs-*/ whereas passive stems are formed using the prefix ተ-/*tə-*/. Infinitive stems are also formed using the prefixመ-/*mə-*/ and reduplicative stems are formed by duplicating the middle consonant. For example, the wordsከተሰበረ/*kətəsəbərə*'if it is broken'/ and ሰባበራችሁ/*səbabəratʃihu*'yourepeatedly broke'/ have the derived stems ተሰበር-/*təsəbər-*/ and ሰባበር-/*səbabər-*/, respectively, but a common basicstemሰበር-/*səbər-*/. From the semantics point of view, there is no conceptual difference between derived and basicstems. Moreover, the derivedstems are generated from basicstems which are common forms for many variants than derivedstems. Therefore, in the case of stem-based indexing and retrieval, variants of a word are represented using their basicstems.

Stemming is usually applied since lemmatization is more computationally consuming for just slight effectiveness improvements (Balakrishnan and Lloyd-Yemoh, 2014). Stemming has also been applied in Amharic IR systems (Mindaye*etal*., 2010; Alemayehu and Willett, 2003).However, stemming is not expected to produce the desired result of term frequency in Amharic text. Morphological variants of Amharic verbs can have more than one stem. For example, morphological variants such as ሰበረ /*səbərə*'he broke'/, ተሰባሪ /*təsəbari*'broken'/, and አሳበረ/*ʔəsabərə*'he helped to break'/ have the basic stems ሰበር- /*səbər-/*, ሰባር- /*səbar-/* and ሳበር- /*sabər-/*, respectively. As a result, stemming provides distorted frequency since each stem of variants is counted differently though they are semantically similar. Therefore, Amharic verbal stems need one more reduction analysis to extract root. Indeed, verbal stems are themselves formed from roots.

4.2.2 Root-BasedMorphological Analysis

Roots are the basis for the formation of basic stems and many other variants of the same Amharic word. Though some words are derived from stems, their origins are roots rather than stems.For example, the morphological structures of the verb ከሰበርኳቸው /*kəsəbərk*^{w}*atʃəw*'if I break them'/,the derived noun ስብራቴ /*sibirate*'my brokenbelonging'/ are as follows.

ከሰበርኳቸው	ስብራቴ

[1] *1*: first person,*2*:second person, *3*: third person, *s*:singular,*pl*:plural,*pre*:preposition,suf:suffix, nom:nominative,*con*:conjunction,*neg*:negative, *gen*:genitive, *def*: definite marker, *adj*: adjectivizer, *sub*: subject, *obj*: object

ከ_ሰበር_ኩ_አቸው	ስብር_አት_ኤ
pre-stem-1,s-3,pl	*stem-nom-1s*
ከ_ስ-ብ-ር_ኩ_አቸው	ስ-ብ-ር_አት_ኤ
pre-root-1,ps-3,pl	*root-nom-1ps*
from-break-I-them	break-nom-my

In the above example, stems of the two words have different forms while they have the same root i.e., ስ-ብ-ር/*s-b-r*/.According to Yimam (2001), more than 10 basicstems can be generatedfrom a given root. As depicted in Table 1, variants of words describe similarconcept, but they have different stems. This affects the term frequency which has an impact on ranking and retrieval of documents. On the other hand, all variants have a single root. Therefore, root-basedrepresentation maps morphologically related words into one common form. Accordingly, statistics information based on root-based approach can be computed accurately so that the actual term frequency can be known.The actual number of variants will be equal to the frequency of their common root. However, the frequency of each stem will be less than the actual occurrence of variants. Moreover, root form increases the matching possibilities between query terms and index terms. Thus, the root-based approach becomes a better way to represent documents and queries for Amharic IR.We have also experimentally analyzed the viabilities of stem-based and root-basedapproaches.

4.3 Stopword Removal

Stopwordsare words thatevenly occurin many documents and serve as purpose rather than content. Thus, they are removed from documents and query.Stopwords can be removed either by applying a list or IDF. In morphologically simple languages like English, stopword identification and removal is achieved by considering a list of words that are identified to be stopwords. The conventional trend applied so far for removing Amharic stopwords is also to use a list. However, taking the characteristics of language into consideration,this is certainly not the most appropriate way. Indeed, Amharic stopwords:(*i*) do notnecessarily exist as standalone words; (*ii*) can accept prefixes and suffixes; and (*iii*) mayexist as part of Amharic words and serve as prefix or suffix. For these reasons, it is not possible to find and remove all Amharic stopwords unless the morphological structure of words is known. For example, one may consider words like "the" a stopword in English. Its Amharic equivalent is a suffix "-ኡ/*-u*/" or "-ው/*-w*/" that does not appear as a standalone word. Accordingly, "the house" and "the student", for instance, are equivalent to ቤቱ /*betu*/(ቤት /*bet*/ + -ኡ/*-u*/)and ተማሪው/*təmariw*/ (ተማሪ/*təmari*/ + -ው/-w/), respectively. As there could be several sequences of affixes representing articles, prepositions, numbers, etc., words can appear in various morphologicalstructures. It means that one could not work with surface forms of words to identify and remove stopwords as most of the stopwords in Amharic do not exist as standalonewords. This indicates that stopword identification and term representation in Amharic IR demands a different consideration than the conventional trend.

Yeshambel*et al.* (2020a) constructedroot-based and stem-based Amharic stopword lists by considering the semantics of Amharic words and corpus statistics. The values of frequency,variance, entropy and mean in a large corpus were used while constructing the stopword list.

Table 1: Amharic root and stems of variants.

Root	Basicstems	Variants	Concept
ት-ል-ቅ	ትልቅ	ትልቁ, ትልቃችን, ትልቃቸው,ትልልቅ, etc.	big
	ታላቅ	ታላቁ, የታላቁ, ከታላቁ, ታላላቆቹ, ታላቅ, etc.	
	ተለቅ	ተለቀ, ተለቀች, ተለቁ, ተለቅን, ተለቃችሁ, etc.	
ም-ስ-ክ-ር	መሰከር	መሰከረ, መሰከረች, መሰከሩ, ተመሰከረ, etc.	witness
	መሳከር	ተመሳከረ, አመሳከረ, ይመሳከራል, etc.	
	መሳክር	እናመሳክር, አመሳክር, አመሳክሪ, etc.	
	መስክር	ይመስክር, እንመስክር, ልመስክር, etc.	
	መሰክር	እንመሰክራለን, ይመሰክራሉ, ትመሰክራለች,etc.	

In both cases,itis shown that stopwords significantly impact on retrieval effectiveness, size of index, and term weighting of non-stopwords. Experimental results also showed that the root-based approach is better than the stem-based approach in conflating all variants of a stopword.The identified stopwordsinclude prepositions (e.g.ወደ/*wədə*'to'/, ስለ/*sɨlə*'about'/, እስከ/*ʔɨskə*'up to', በ-/*bə*-'by'/, ከ-/*kə*-'from'/, etc.), conjunctions (e.g.እና/*ʔɨna*'and'/, ያሁንእንጂ/*yɨhunʔɨndʒɨ*'however'/, እዚህ/*ʔɨzih*'here'/, etc.), negation markers (አል…ም/*ʔəl…m*'not'/), indefinite articles (አንድ/*ʔənd*'an'/), auxiliary verbs (ኦ-ል/*ʔ-l*'say'/, ን-በ-ር/*n-b-r*'was'/, etc.), ወዘተ/*wəzətə*'and so on'/, etc.

4.4 Indexing

To test the effect of morphological analysis on Amharic IR,stem-based and root-based indexes are created using Lemur toolkit. Lemur is also used for retrieval purpose. The stem-based index was created using the basicstems of words while the root-based index was created using the root of words.

4.5 Matching

In the proposed system, document processing involves text preprocessing, morphological analysis, stopword removal and indexing. As a result of this process, we obtain indexed documents. On the query side, we apply similar processes except indexing. Thus, query processing provides a set of terms representing information needs of users. Searching of relevant documents is carried out by matching query terms (representing information need of users) with index terms (representing documents). In this work, we use exact vocabulary term matching. It searches documents that contain the query terms without analysing the semantics of words and without considering the semantic connections between them.

4.6 Ranking

For a given query Q and the collection of retrieved documents D, the Lemur toolkit ranked retrieval results based on their relevance. The document length and number of matching query terms are considered. OKAPI ranks documents based on the following algorithm.

$$\text{Score (D,Q)}=\sum_{i=1}^{n} IDF(\text{qi}).\frac{f(qi,D).(k1+1)}{f(qi,D)+k1.(1-b+b.\frac{|D|}{avgdl})} \quad (1)$$

Where f(qi,D) is qi's term frequency in the document D, |D| is the length of the document D in words, and avgdl is the average document length in the text collection from which documents are drawn. k1 and b are free parameters. In this research the value of k1 = 2.0 and b = 0.75. IDF(qi) is the IDF (inverse document frequency) weight of the query term qi.

In case of LM, the KL divergence ranking function with default parameters was used. It captures the term occurrence distributions and computed as follow.

$$\sum_{w:c(w,d)>0,p(w|\theta Q)>0} p(w|\theta Q) log \frac{ps(w|d)}{\alpha d \text{p(w|C)}} + log \alpha d \quad (2)$$

Where d is document, w is word, p(w|θQ) is a query model, is estimatedquery, p(w|C) is the collection language model, ps(w|d) is the smoothedprobability of a word seen in the document, αd is a coefficient controlling.

5 EXPERIMENT

5.1 Experimental Data

The experimental setup followed the format of the Text Retrieval Conference (TREC). Experiments were carried out using a scientifically built corpus(Yeshambel*et al*., 2020b) and a stopword list constructed by Yeshambel*et al*. (2020a). The test collection has 12,538 documents and 240 queries while the stopword list contains 222 stopwords.

5.2 Implementation and Measures

Python was used for the preprocessing tasks while indexing and retrieval were performed using Lemur toolkit, which is a search engine designed to support research on language model (LM) for IR tasks[2]. The retrieval effectiveness was evaluated automatically using trec_eval tool which can compute many evaluation measures[3]. LM and BM25 models were used as retrieval models.

[2]http://www.lemurproject.org
[3]http://trec.nist.gov/trec_eval

5.3 Results and Discussion

5.3.1 Results

To investigate the effectiveness of Amharic system with respect to various word forms used for document representation, we conducted three retrievalexperiments: word-based, stem-based, and root-based retrieval. The retrieval effectiveness is shown is Table 2.

Table 2: Retrieval effectiveness based on the three approaches.

Approach	Precision				
	P@5	P@10	P@15	P@20	MAP
Word	0.56	0.49	0.44	0.40	0.43
Stem	0.62	0.53	0.47	0.43	0.57
Root	0.79	0.70	0.61	0.55	0.70

The root-based approach retrieves more relevantdocuments than stem-based and word-basedapproaches. It has also rejected non-relevant documents better than stem-based and word-based approaches. The word-based and stem-based methods miss more relevant documents since they cannot handle some morphological variations. The retrieval effectiveness of the three approaches decreases from precision @5 documents to precision @20 due to scarcity of relevant documents in the test collection.

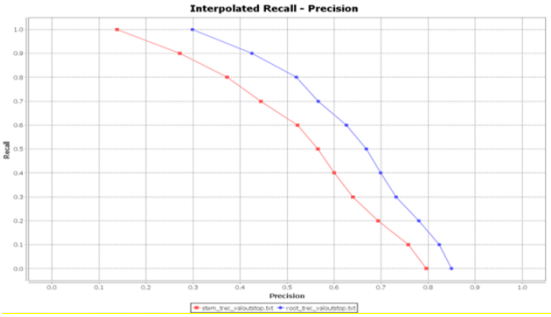

Figure 2: Recall-precision-curveof stem and root.

The recall and precision values of stem-based and root-based approaches are shown in Figure 2. The blue line depicts the root-based retrieval effectiveness, whereas the red line represents the stem-based retrieval results. It can seen that the retrieval effectiveness of root-based approach outperforms stem-based approach.

5.3.2 Discussion

Comparison of root and stem for retrieval

Although the stem-based approach could not conflate all variants, it improves retrieval effectiveness to some extent. However, it affects the actual term frequency of some word classes which results in loss of the rank of retrieved relevant documents.Some relevant documents which are not retrieved withstem-based approach are retrieved using root-basedretrieval. Furthermore, some non-relevant documents retrieved in the case of stem-based approach are not retrieved with root-basedapproach. There are three reasons behind this.

First, root can conflate all morphologically variants to one common form, but not stem. For example, the stems of variants ሰበረ/*səbərə*/, ስበር/*sɨbər*/, ስብራት/*sɨbɨrat*/,አሳበረ/*ʔəsabərə*/, ሰባራ/*səbara*/, አሳብራት/*ʔəsabrat*/, ተሰብሮ/*təsəbro*/ are ሰበር-/*səbər-*/, ስበር-/*sɨbər-*/, ስብር-/*sɨbɨr*/, ሳባር-/*sabar-*/, ሰባር-/*səbar*/, ሳብር-/*sabr-*/, and ስብር-/*səbr-*/, respectively. This creates termmismatch with each other. However, all variants have one common root ስ-ብ-ር/*s-b-r*/.Therefore, the root-based representation increases the term frequency which usually leads to better retrieval result.

Second, root forms donot conflate semantically unrelated words to a common form. However, the stem-based approach sometimes conflates semantically unrelated words. For example, ገደል/*gədəl*/ is the stem of the verb ገደለ/*gədələ*'he killed'/ and the noun ገደሎች/*gədəloʧ*'cracks'/. However, their roots are ግ-ድ-ል /*g-d-l*/ andገደል/*gədəl*/,respectively.The verb ሲገድል/*sigədɨl*'as he kills'/ and the noun ገድል/*gədɨl*'contending'/ have the same stem ገድል/*gədɨl*/. However, their roots are ግ-ድ-ል /*g-d-l*/ and ገድል/*gədɨl*/, respectively. This indicates that the use of stemsleads to retrieval of non-relevant documents. Therefore, the stem-based approach is not powerful to filter out non-relevant documents.

Third, the retrieval result of the stem-based approach depends on the query word variants while this is not the case with the root-basedapproach. The root-based approach performs equally for all the variants of the query terms. However, the stem-based approach returns different results in different ranks. In Amharic, different users will certainly construct the same information need using different word variants. For example, the query 'the causes of air pollution' can be constructed as:

- የአየርንብረትብክለትመንስኤዎች/*jəʔəjərnɨbrətbɨklət mənsiʔewəʧ*/;

- ለአየርንብረትመበከልመንስኤወች/*lə?əjərnɨbrətməbə kəlmənsɨ?ewəʧ/*;
- የአየርንብረትበካይመንስኤወች/*jə?əjərnɨbrətbəkajm ənsɨ?ewəʧ/*; etc.

After the stem-based morphological analysis, all the three queries will have same query terms (አየር, ንብረት and መንስኤ) except one (ብክል, በክል and በካይ). As a result of variation of the thirdterm, the system returns different results in different ranks. Therefore, stem-based approach performs differently for the same test collection (see Figure 3). The top line (in green) depicts root-basedretrieval whereas the remaining two lines representstem-based retrieval.**Comparison with previous studies**

In the previous studies, a few Amharic IR systemshave been evaluated. Some of them are based on stems (Mindaye*et al*., 2010; Munye and Atnafu, 2012) while some others are based on citation forms (Argaw*et al*., 2004) and root-based (Alemayehu and Willett, 2003). However, due to the complexity of the language the stem-based and n-gram models do not work well. In this work, we have shown that the rootsaremore powerful for Amharic IR than stems. This is a new finding which was not looked in the previous researches. Other authors suggestedstem-based as the best option. Alemayehu and Willett (2003) investigated the effects of the stem and root-based approaches on Amharic IR. Their finding was that the stem-based approach is better than the root-based one. The justification was that many Amharic words have common root though they are semantically unrelated. Their report states that stem-based improves the root-based method in terms of rejecting irrelevant documents that could be retrieved by the use of roots. However, their experiment was carried out on wrong representation of roots. For example, the root of the word ዝናብ/*zɨnab* 'rain'/ is misrepresented asዝንብwhich is the same as ዝንብ /*zɨnb*'fly'/. However, the correct root representation is ዝ-ን-ብ /*z-n-b*/. Another problem in their approach is that they used root representation for any types of words by removing vowels in non-derived words as well. This method conflates many semantically unrelated words. Furthermore, their system wrongly extracts the root radicals of some words. For example,ምትis considered asthe root of the words ሞተ/*motə*'die'/ and መታ/*məta*'hit'/. But, their roots are ምˆ-ት and ም-ት, respectively. Such cases are prevalent in Amharic. Thus, previous studies that recommended the use of stems made their conclusions without through investigation on the applicability of roots.

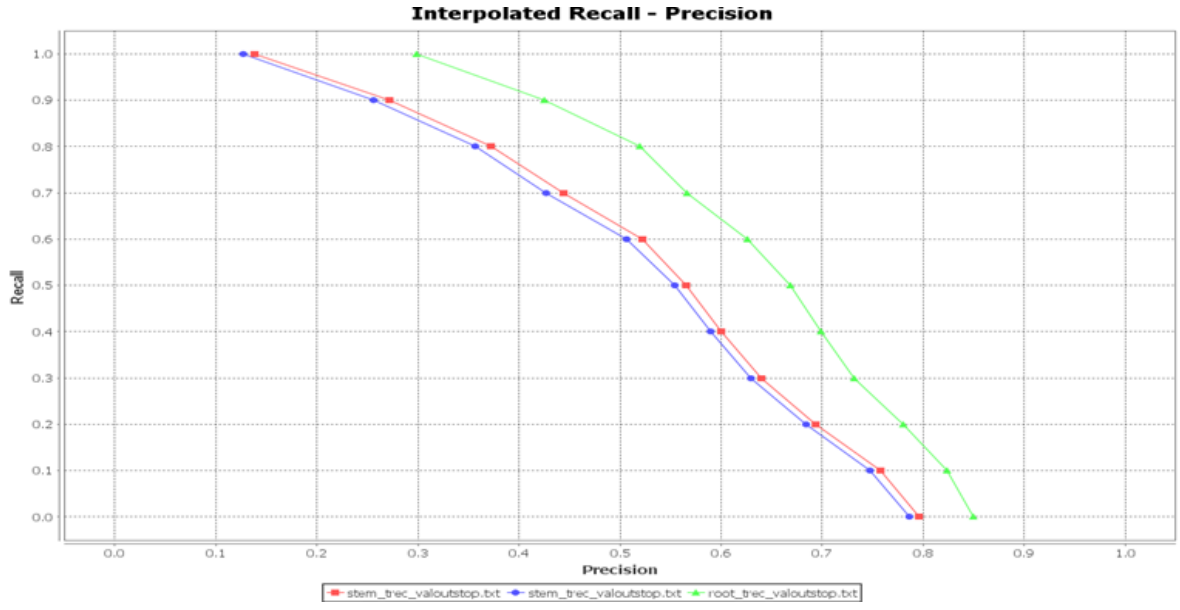

Figure 3: Recall-Precision curves based onstemsand roots.

Comparison with Google Amharic retrieval

The Google Amharic search engine is based on stem. It returns different retrieval results in different ranks for the same query using different variants of query words. Similar results are obtained inour work while we apply stem-based approach. For example, Google search results of the queries የአጥንትስብራት/*jəʔət'ɨntsɨbɨrat*'being broken bone'/ and የአጥንትመሰበር/*jəʔət'ɨntməsəbər*'the process of being broken bone'/ are different though the same concept is expressed via different variants. Our approach differs from Google search engine into two ways.

i. Google searches based on both basicstems and derived stems. It returns different retrieval results for basic stems and derived stems queries though they are semantically similar. However, in our work, the stem-based approach is based on basic stems only, providing the same retrieval results for both basic stems and derived stems.

ii. Google does not employ roots to represent verbs and words derived from them. However, we use root-based approach as itconflates all variants of words to a common form.

Comparison of LM and BM25

We also compared the performance of LM and BM25. Although LM is very popular and powerful for IR in different languages, previous Amharic IR researches were made based on classical IR models such as vector space model(Mindaye*et al*. 2010;Argaw*et al*., 2004; Argaw*et al*., 2006). Language modelling was not employed for Amharic IR. In this study, the impact of language modelling retrieval on Amharic IR is also investigated.As depicted in Figure 4,the blue line representing LM is above the red line representing BM25. Both precision and recall values of LM are better than that of BM25 at different levels. This is because of the capability of LM to capture the dependency of words and estimate the probability distribution of a query in each document. This means LM is a more suitable retrieval model for Amharic language.The roots of verbs and words derived from them and the basic-stems of other word classes are robust to represent words not only in IRbut also in other applications such as machine translation, information extraction, sentiment analysis, etc.

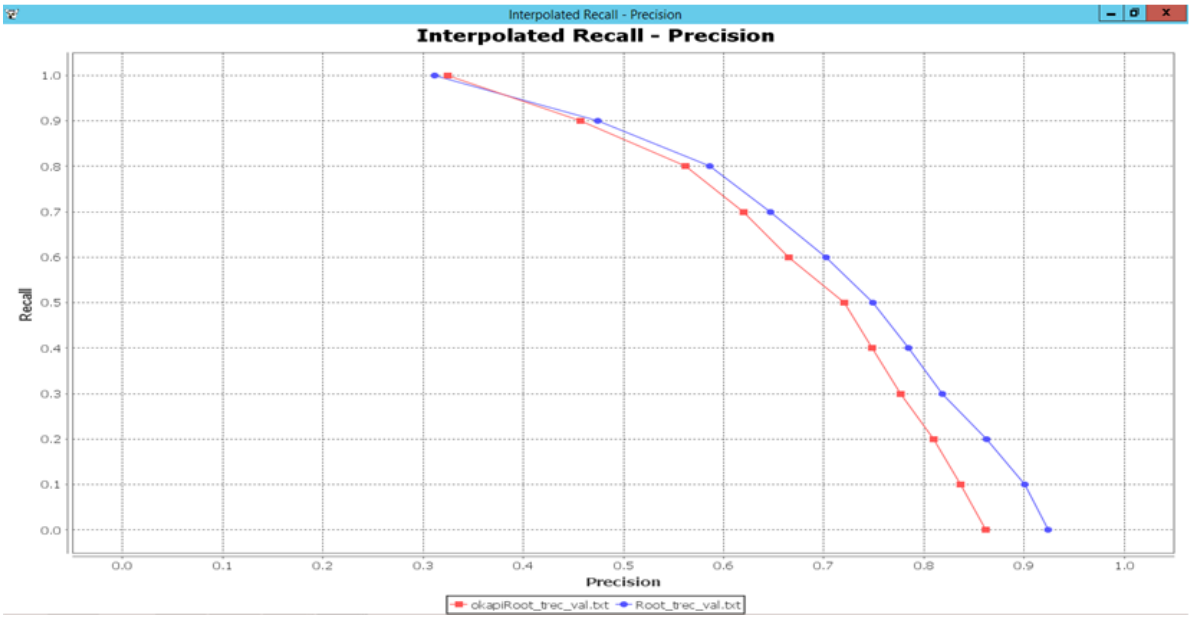

Figure 4: Comparison of LM and BM25 retrieval models on root-based approach.

6 CONCLUSION

The development of Amharic IR demands thorough investigation of the characteristics of the language. Its complex morphology affects the way documents and queries are represented for the task of information retrieval. In this work, we conducted several experiments using various forms of words for document representation. Experimental results have shown that root forms of words provide better results in representing documents. It is also shown that, with the use of root forms, the LM retrieval is better than BM25 model. As the proposed system is based on exact vocabulary term matching, future work needs to consider query expansion so as to take into account synonyms, collocation words, name identification, etc.Furthermore, the language has many ambiguous words which have different meaning in various contexts. Thus, this work may be improved by handling ambiguity.

REFERENCES

Abate, M. and Assabie, Y., 2014. Development of Amharic morphological analyzer using memorybased learning. In *Proc. of the 9th Int. Conf. on Natural Language Processing*, pp. 1-13, Warsaw.

Abdusalam, A.,2008. Effective retrievaltechniques for Arabictext. *PhD Thesis*, School of Computer Science and Information Technology, RMIT University, Melbourne, Victoria, Australia.

Alemayehu, N. and Willett, P., 2003. The effectiveness of stemming for information retrieval in Amharic.*Program: Electronic Library and Information Systems*, 37(4), pp.254–259.

Al-Hadid, I., Afaneh, S., Al-Tarawneh, H., and Al-Malahmeh, H., 2014.Arabic information retrieval system using the neural network model.*International Journal of Advanced Research in Computer and Communication Engineering. Vol. 3, Issue 12.*

Ali, A., Mosa, E., and Abdullah, B., 2020.An intelligent use of stemmer and morphology analysis for Arabic information retrieval.*Egyptian Informatics Journal*, https://doi.org/10.1016/j.eij.2020.02.004.

Ambati, V., Rohini, U., Pramod, P., Balakrishnan, N., and Reddy, R., 2006. Multilingual information access : Information Retrieval and Translation in a Digital Library. In: Proc. of the 2nd Int. Conf. on Universal Digital Library, Alexandria, Egypt.

Argaw, A. A., Asker, L., Cöster, R., and Karlgren, J., 2004. Dictionary-based Amharic-English information retrieval.In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pp. 143–149, Springer.

Argaw, A. A., Asker, L., Cöster, R., Karlgren, J., and Sahlgren, M., 2006. Dictionary-based Amharic-French information retrieval. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4022 LNCS*, 83–92. https://doi.org/10.1007/11878773_9.

Assabie, Y., 2017. Development of Amharic morphological analyzer.*Technical Report*.Ethiopian Ministry of Communication and Information Technology, Addis Ababa, Ethiopia.

Balakrishnan, V., and Lloyd-Yemoh, E., 2014. Stemming and Lemmatization: A Comparison of Retrieval Performances. *Lecture Notes on Software Engineering*, *2*(3), 262–267. https://doi.org/10.7763/lnse.2014.v2.134.

Cambria, E. and White, B., 2014. Jumping NLP curves: A review of natural language processing research, *IEEE Computational Intelligence Magazine*, 9(2): 48-57.

Carmel, D. and Maarek, Y., 1999.Morphological disambiguation for Hebrew search systems. *International Workshop on Next Generation Information Technologies and Systems;NGITS 1999: Next Generation Information Technologies and Systems*, pp. 312-325.

Darwish, K. andMagdy, W., 2014.Arabic information retrieval.*Foundations and Trends in Information Retrieval*, 7(4): 239-342.

Jackson, P. and Moulinier I.,2007. Natural language processing for online applications: *Text retrieval, extraction and categorization*. John Benjamins Publishing, Amsterdam,Netherlands, 2nd ed.

Larkey, L. S., Ballesteros, L., and Connell, M. E. 2007. Light stemming for Arabic information retrieval. In *Arabic Computational Morphology*,pp.221-243. Springer, Dordrecht.

Mindaye, T., Redewan, H. and Atnafu, S., 2010. Design and implementation of Amharic search engine. In *Proc. of the 5th Int. Conf. on Signal Image Technology and Internet Based Systems*, pp. 318–325.

Munye, M. and Atnafu, S., 2012. Amharic-English bilingual web search engine. *Proceedings of the International Conference on Management of Emergent Digital EcoSystems, MEDES 2012*,pp. 32–39.

Moukdad,H., 2002. A comparison of root and stemming techniques for the retrieval of Arabic documents.

PhD Thesis, Graduate School of Library and InformationStudies, McGill University, Montreal.

Musaid, S., 2000.Arabic information retrieval system-based on morphological analysis (AIRSMA): *A* comparative study of word, stem, root and morpho-semantic methods. *PhD Thesis*,Computer and Information Science, DeMontfort University, United Kingdom.

Ornan, U.,2002. A morphological, syntactic and semantic search engine for Hebrew texts.In:*Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages*.Philadelphia, Pennsylvania, USA.

Sanderson, M., and Croft, W., 2012.The history of information retrieval research.In*Proceedings of the IEEE*, Special Centennial Issue, pp.1444-1451.

Smeaton, A. F., 1992. Progress in the application of natural language processing to information retrieval tasks.*The Computer Journal*, 35(3), pp. 268-278.

Yeshambel, T., Mothe, J. and Assabie, Y.,2020a.Construction of morpheme-based Amharic stopwordlist for information retrieval system.In: *Proc. ofICAST2020*,Springer.

Yeshambel, T., Mothe,J. and Assabie, Y., 2020b.2AIRTC: The Amharic Adhocinformation retrieval test collection. In:*Proc. ofCLEF2020*, LNCS 12260, pp. 55-66.

Yimam, B., 2001. Yamarignasewasiw (Amharic grammar).CASE.Addis Ababa, 2nd edition.