



HAL
open science

Amharic Semantic Information Retrieval System

Tilahun Yeshambel, Josiane Mothe, Yaregal Assabie

► **To cite this version:**

Tilahun Yeshambel, Josiane Mothe, Yaregal Assabie. Amharic Semantic Information Retrieval System. 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2020), Nov 2020, virtual, France. pp.22-44, 10.1007/978-3-031-14602-2_2 . hal-03855982

HAL Id: hal-03855982

<https://ut3-toulouseinp.hal.science/hal-03855982v1>

Submitted on 16 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Amharic Semantic Information Retrieval System

Tilahun Yeshambel¹(✉), Josiane Mothe², and Yaregal Assabie³

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. Amharic is the official language of Ethiopia, currently having a population of over 118 million. Developing effective information retrieval (IR) system for Amharic has been a challenging task due to limited resources coupled with complex morphology of the language. This paper presents the development of Amharic semantic IR system using query expansion based on deep neural learning model and WordNet. In order to optimize the retrieval result, we propose Amharic text representation using root forms of words applied for stopword identification, indexing, term matching and query expansion. Comparisons are made with the conventional stem-based text representation for information retrieval, and we show that using the root forms of words is better for both resource construction and system development. The effectiveness of the proposed Amharic semantic IR system is evaluated on Amharic *Adhoc* Information Retrieval Test Collection (2AIRTIC).

Keywords: Semantic information retrieval · Query expansion · Complex morphology · Amharic IR resources

1 Introduction

Searching information on a huge corpus is one of the common tasks nowadays. Information Retrieval (IR) focuses mainly on the process of matching user queries terms to index terms in order to locate relevant documents from Web or corpus. Matching query terms with index terms is one of the main challenges of IR in many languages [1]. Linguistic variation of a natural language and term mismatch affects the effectiveness of IR system. As a result of linguistic variation, some relevant documents for a user need will be omitted from a search result. Natural language processing (NLP) has significant role in many languages IR systems for extracting index and query terms. It is applicable to reduce the space required for indexing and maximizing the retrieval effectiveness by conflating variants of words to a common form [2]. Stemming is one of the typical NLP techniques to handle morphological variants in many languages and is still an active research topic specifically for under resourced languages [3, 4]. Furthermore, NLP techniques are employed to handle term matching for semantically related concepts.

Semantic text matching in IR is the task of finding semantic similarity between query and document text. Sometimes, since a user information need is imprecise, incomplete and semantically ambiguous, a retrieval system cannot retrieve relevant documents to a query. Query expansion is the task of adding semantically related terms to a given query for improving the performance of IR system. Different approaches have been proposed for expanding query terms such as relevance feedback [5], the use of Wikipedia [6], or resources like thesauri or WordNet [7], and neural network (NN) to capture semantic relationships between words [8].

Research and development on Amharic IR lags behind because of morphological complexity of the language, lack of usable NLP tools, resources, and test collection. Despite many works on IR for many languages, few researches have been conducted on Amharic IR. The existing Amharic IR systems face challenges in searching relevant documents because of the morphological complexity and semantic richness of the language. Amharic exhibits complex morphology that poses challenges in NLP and IR [9, 10]. The base of Amharic word can be stem or root. The morphological structure (root or stem) one should choose for indexing, matching, and resource construction is an open question in Amharic IR. Relevant documents for an Amharic query may not be retrieved as a result of term mismatch between index terms and query terms. Amharic stem-based indexing and term matching misses some relevant documents because multiple stems exist for variants. For example, the variant ፈለገ /*fəlagə*/, ፈለገ /*fəlagi*/, ፍለጋ /*filəga*/, እንፈልግ /*ʔinifəligi*/, and አፋለገ /*ʔafəlagə*/ have the stems ፈለግ /*fəlagi*/, ፈለግ /*fəlagi*/, ፍለግ /*filəgi*/, ፈልግ /*fəligi*/, and ፋለግ /*fəlagi*/, respectively. On the other hand, a document can possibly be relevant to a user query even if they share semantically similar terms that may even be different variants. For example, ህብረት /*hibirəti* ‘union’/, አንድነት /*ʔanidinəti* ‘unity’/, አብሮነት /*ʔəbironəti* ‘fellowship’/, ቅንጅት /*k’inidziti* ‘coalition’/, ትብብር /*tibibiri* ‘cooperation’/ and ጥምረት /*timirəti* ‘collaboration’/ are semantically related words. Both cases, i.e. same variants with different stems and similar semantics with different variants, lead to poor IR performance as relevant documents are missed. Therefore, Amharic retrieval system needs to identify optimal representative of variants and reformulate a user query by expanding initial user query in order to retrieve more relevant documents. In this paper, we investigate the characteristics of the language and suggest root forms of words to handle morphological variations to increase the quality of indexing and the probability of matching between index and query terms. Furthermore, we investigate the impact of query expansion on Amharic semantic IR using word embedding and WordNet.

The rest of this paper is organized as follows. Section 2 briefly describes the characteristics of Amharic language. Section 3 discusses related work. Section 4 presents the design of Amharic semantic IR system whereas Sect. 5 presents Amharic resources that we constructed. Experimental results are discussed in Sect. 6. Finally, conclusion and future research directions are forwarded in Sect. 7.

2 Amharic Language

Amharic is the official language of Ethiopia that has a population of over 118 million at present [11]. It has been used as a working language of the government of the country for

a long time. As a result, its rich literary heritage has endowed the language with huge written resources and it serves as a *lingua franca* of the country. Amharic belongs to Semitic language families and has its own script which has alphabet, numbers and punctuations. The alphabet has 33 basic characters and each of them has 7 different forms representing consonant-vowel combination. The vowels are ኧ /ʔə/, ከ /ʔu/, ኢ /ʔi/, ኣ /ʔa/, ኤ /ʔe/, ኦ /ʔi/, and ኦ /ʔo/. For example, consonant-vowel combination of the base character ከ /kə/ has the following modifications: ኩ /ku/, ኪ /ki/, ካ /ka/, ኬ /ke/, ክ /ki/, and ኮ /ko/. Furthermore, the alphabet has labialized characters such as ሷ /lʷa/, ሸ /mʷa/, ሹ /sʷa/, and ቈ /qʷa/. Their structure is consonant-vowel-vowel combinations.

Amharic is morphologically rich and complex agglutinative language. It is considered as one of the most prolific languages [9]. Clitics such as prepositions, articles, conjunctions, and pronouns are glued to nouns, verbs, and adjectives. The internal structure of a word may include the base (i.e. stem or root), affixes, and patterns. Amharic word may contain many affixes which are attached in complex rules. An Amharic word can represent a sentence in another language. For example, ሰበረኞቻቸው /səbərətʃatʃəwi/ 'she broke them' is an agglutination of the verbal stem (ሰበር /səbəri/ 'broke'), subject marker pronoun (ኧች /ʔətʃi/), and object marker pronoun (ኣቸው /ʔətʃəwi/). Amharic words can be classified as *derived* and *non-derived*. Derived words are formed from other word classes through derivational process. The word formation in both cases usually involves change in one or more characters of a stem or root. The change arises as a result of making a word for case, gender, number, tense, person, mood, etc.

3 Related Work

3.1 Conventional Information Retrieval

Although Amharic is widely used in Ethiopia, the status of IR system development for the language is relatively at rudimentary level. The retrieval effectiveness of stem-based and root-based text representations on Amharic language are studied in [12]. Experiments were carried out by running 40 queries on 548 documents using OKAPI system and the study concluded that stem-based retrieval is slightly better than root-based one. Amharic search engine was developed using stems and tested by running 11 queries on 75 news documents [13]. The average precision and recall values were 0.65 and 0.95, respectively using OR operator for query terms, and 0.99 and 0.52, respectively using AND operator. Arabic is a Semitic language for which relatively more IR research is conducted. For example, Al-Hadid *et al.* [14] developed a neural network-based model where documents and queries are represented using stems and their similarity is computed using cosine similarity. The effectiveness of Arabic word-based, stem-based, and root-based representation of documents and queries was investigated by Musaid [15]. The word-based and stem-based representations miss relevant documents while root-based one retrieves non-relevant documents. The effects of stem and root on Arabic search engine was also compared by Moukdad [16]. The results indicated that stemming is more effective than root. A comparison between stem-based and root-based Arabic retrieval was made by Larkey *et al.* [17]. The finding indicates that light stemmer outperforms root analyzer and other stemmers which are based on detailed morphological analysis. Ali *et al.* [18] 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 results showed slight improvement on IR effectiveness due to the stemmer. Hebrew is one of the Semitic languages spoken mainly in Israel. Ornan [19] designed Hebrew search engine by applying a rule-based morphological analysis. The design of the search engine takes into account the construction of a morphological, syntactic and semantics analyzer. Words unsuited for the syntax and the semantic of a sentence were removed.

3.2 Semantic Information Retrieval

Amharic semantic-based IR using BM25 was developed [20]. Documents and queries were processed by a stemmer. The system was evaluated by running 10 queries on 8,759 documents and performed an average recall and precision of 0.84 and 0.23, respectively. Fang [21] developed and evaluated English semantic retrieval system using WordNet and dependency-thesaurus on TREC test collections. Query terms are expanded considering term relationships and synset definition of a word in the WordNet and mutual information in the collection. The retrieval results indicated that significant improvement was achieved after query expansions using both methods. Better retrieval result was obtained using synset definition of terms. Retrieval based on thesaurus is less effective than definition-based retrieval. The impact of integrating word embedding and entity embedding with and without interpolation within the *ad hoc* document retrieval task was studied and evaluated on TREC collections (ClueWeb'09B and 100 ClueWeb'12B) [22]. The authors reported that word embedding do not show competitive performance to any of the baselines (relevance model, sequential dependence model and entity query feature expansion) even after interpolation. CBOW method showed better performance than skip-gram for the *ad hoc* document retrieval task. Entity-based embedding performed better than word-based embedding.

3.3 Evaluation of Amharic IR Corpora, Resources and NLP Tools

Few studies have been conducted to develop NLP tools and create Amharic corpora resources although IR test collections are required for automatic evaluation of IR system. We can quote a few such studies. Demeke and Getachew [23] created Walta Information Center news corpus; Yeshambel *et al.* [24] built 2AIRTC; and Yeshambel *et al.* [10] created stem-based and root-based morphologically annotated Amharic corpora semi-automatically. The sizes of corpora created by Demeke and Getachew [23], Yeshambel *et al.* [24] and Yeshambel *et al.* [10] are 1,065, 12,586, and 6,069 documents, respectively. Mindaye *et al.* [13] and Samuel and Bjorn [25] created Amharic word-based stopword list whereas Alemayehu and Willett [26] built stem-based stopwords list. NLP tools such as stemmer and morphological analyzer have crucial role for processing text documents and user information need. Alemayehu and Willett [26] and Alemu and Asker [27] developed rule-based Amharic stemmers. Sisay and Haller [28] and Amsalu and Gibbon [29] developed Amharic morphological analyzers using Xerox Finite State Tools (XFST) method. Gasser [30] developed rule-based morphological analyzer for Amharic, Tigrinya and Afaan Oromo languages. Mulugeta and Gasser [31] also developed a morphological analyzer using supervised machine learning approach whereas Abate and Assabie [9]

developed morphological analyzer using memory-based supervised machine learning approach.

In this work, we assess the accessibility, quality, and usability of the existing accessible Amharic IR corpora, resources and tools with the purpose of highlighting the status of Amharic language processing applications. The majority of them are not accessible and have limited functionality and size. They are also inconvenient to use or to integrate in Amharic IR experiments. The existing Amharic NLP tools are not full-fledged systems as they are under prototype stages. For example, the stemmer developed by Alemayehu and Willett [26] and Gasser’s morphological analyzer [30] over-stem and under-stem many words. Moreover, these NLP tools extract basic stems of some words and derived-stems of other words. We tested them using a dataset that contains 200 words from different word classes. The stemmer and the analyzer performed 41.4% and 47.6%, respectively. From our experiments, we observed that the performance of Amharic NLP tools on verbs is less than on other word types. Many of the existing test collections are simply sets of documents without topics and relevance judgment. Furthermore, they are small in size compared to test collections created for other languages. Consequently, they would not be used for accurately testing the performance of IR techniques. In our previous work [24], we developed an Amharic IR test collection that consists in a corpus, topic set and the associated relevance judgment. It allows researchers to evaluate retrieval system automatically though the size is still small relative to standard test collections. The test collection is accessible freely at <https://www.irit.fr/AmharicResources/>.

4 Design of Amharic Semantic IR System

Considering the morphological characteristics of the language, we propose a design for Amharic semantic IR system. In the proposed design, the morphological analysis is carried out before stopword removal (see Fig. 1). Morphological analysis is among the key tasks in our IR system as it helps to select index and query terms from documents and queries.

4.1 Preprocessing

Preprocessing includes character normalization as well as tag removal and punctuation mark removal. Character normalization is made to represent various characters having similar pronunciation using a single grapheme. The characters ሐ /hə/, ኅ /hə/, ኆ /hə/ and their modifications are normalized to their corresponding modifications of ሀ /hə/. The character ሠ /sə/ and its modifications are normalized to their corresponding modifications of ሰ /sə/. The character ፀ /ts’ə/ and its modifications are normalized to their corresponding modifications of ጸ /ts’ə/. The character ባ /ʔə/ and its modifications are normalized to their corresponding modifications of ኣ /ʔə/. The fourth orders ሃ /ha/, ላ /ha/, ኃ /ha/ and ኄ /ha/ are normalized to ሀ /hə/ whereas the fourth orders ኣ /ʔa/ and ዓ /ʔa/ are normalized to ኣ /ʔə/. After character normalization, we segment sentences and tokenize words. Sentence segmentation is carried out using punctuation marks used for marking sentence boundaries whereas word tokenization is performed using space, tags and punctuation marks. Tags and punctuation marks are removed after sentence segmentation and tokenization.

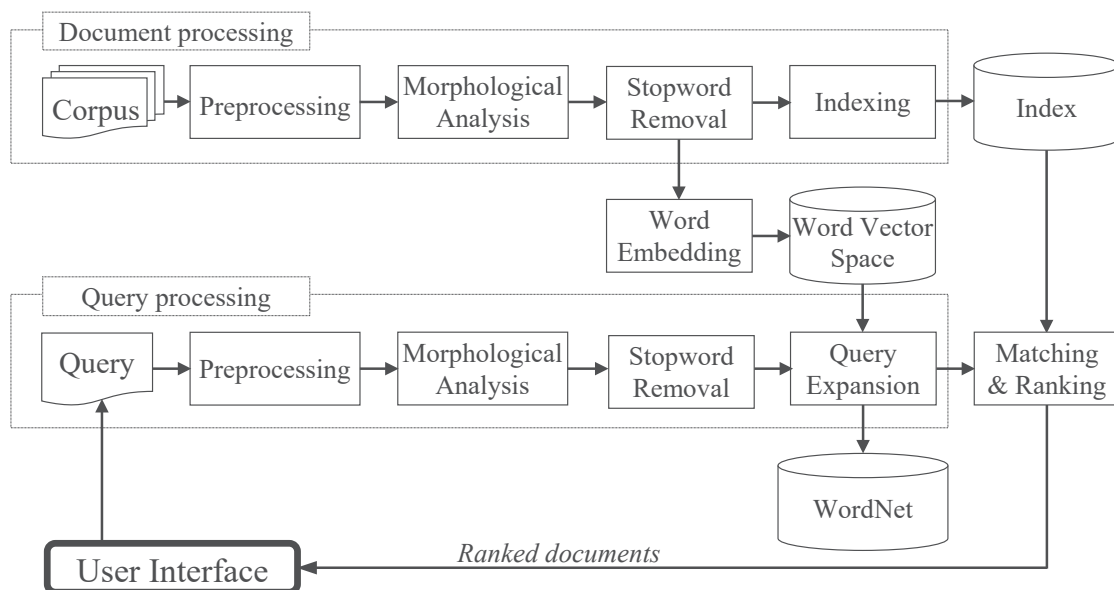


Fig. 1. Design of Amharic semantic IR system.

4.2 Morphological Analysis

Documents and user information need should be represented appropriately using terms that will be used later for matching query with documents. It is to be noted that indexing terms are weighted based on the word 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. Therefore, in this research, we study the feasibility of stem-based and root-based document representation with respect to their effectiveness for Amharic IR. Since well-designed Amharic stemmer and morphological analyzer are not available yet, we design a semi-automatic morphological processor to segment words into their morphemes so that the base of words (i.e. stem and root) could be extracted easily and quickly. The two morphological analyses performed in this work are stem-based and root-based morphological analysis using lexicons created by Yeshambel *et al.* [10]. The lexicons are constructed from a corpus. The stem-based morphological process segments stem of a word from the rest of morphemes whereas the root-based morphological process segments root from the rest of morphemes of a word. For example, the stem-based and root-based morphological segmentation of the verb ስለታወቃቸው /silətawək'atfəwi and the noun ከሰዎች /kəsəwotfu are presented in Table 1. The morphological annotation of different word classes is further presented in Sect. 5.1.

In Amharic, roots are the base of stems. Multiple verbal stems can be generated from an Amharic verbal root (see Table 2). The stems are generated by using different patterns that insert different vowels between root radicals. However, variants of words that are not derived from verbal roots have only one stem which has the same representation as its root. The stems and roots of words are extracted from stem-based and root-based morphologically annotated corpora, respectively, using Algorithm 1.

Table 1. Sample stem-based and root-based morphological segmentation.

Word	Stem-based segmentation	Root-based segmentation
ስለታወቃቸው	ስለ_ተ_አወቅ_አቸው silā_tə_ṗawək'i_ṗətfəwi [pre ¹]-[pas]-[stem]-[3,pl]	ስለ_ተ_አ-ው-ቅ_አቸው silā_tə_ṗ-w-k'_ṗətfəwi [pre]-[pas]-[root]-[3,pl]
ከሰዎቹ	ከ_ሰው_አቸ_ኡ kə_səwi_ṗotfi_ṗu [pre]-[stem]-[pl]-[def]	ከ_ሰው_አቸ_ኡ kə_səw_ṗotfi_ṗu [pre]-[root]-[pl]-[def]

¹ *l*: first person, *3*: third person *s*: singular, *p*: plural, *f*: feminine, *pre*: preposition, *loc*: focus, *pas*: passive, *nom*: nominative, *conj*: conjunction, *neg*: negative, *gen*: genitive, *def*: definite marker, *adj*: adjectivizer, *pos*: possessive, *acc*: accusative, *pal*: palatalizer, *comp*: complement.

Algorithm 1. Extracting stem and root from corpus.

Input : Affix lists and Annotated corpora
Output : Stem-based and root-based Corpora
Step 1: Open affix lists and annotated corpora
Step 2: For each document in the annotated corpus:
For each word in a document:
Segment a word into morphemes using '_'
If a morpheme is in affix lists
Delete from an annotated document
End if
End for
End for

4.3 Stopword Removal

Stopwords are words that evenly occur in many documents and serve as purpose rather than content. Thus, as they are non-content bearing terms, they are removed from documents and queries in IR systems. As shown in our proposed design (Fig. 1) Amharic stopwords are removed after morphological analysis is carried out on documents and queries, which is different from the design of IR for morphologically simple languages. In morphologically simple languages like English, stopword identification and removal is made before stemming by using stopword list. The conventional trend applied so far for removing Amharic stopwords is also to use a stopword list, and it is carried out before stemming or morphological analysis. However, taking the characteristics of the language into consideration, this is certainly not the most appropriate way. Indeed, Amharic stopwords are characterized by the following three morphological features: (i) they do not necessarily exist as standalone words; (ii) they can accept prefixes and suffixes; and (iii) they may exist 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

Table 2. Sample of Amharic verbal roots and basic stems of variants.

Root	Stem	Variant	Concept
ጥ-ቅ-ም	ጠቀም	ጠቀመ, ጠቀመኝ, ሲጠቀም, እንዲጠቀም, etc.	Use
	ጠቃም	ጠቃሚ, የጠቃሚው, ተጠቃሚ, ጠቃሚያችን, etc.	
	ጠቅም	ሲጠቅም, ይጠቅም, ሊጠቅም, ጠቅሞኛል, etc.	
	ጥቅም	ጥቅሙን, ከጥቅሜ, ለጥቅማቸው, etc.	
	ጥቀም	ጥቀሙ, መጥቀሙ, እንጥቀም, etc.	
ም-ር-ም-ር	መረመር	መረመረ, መረመሩ, ስመረመር, ተመረመረች, etc.	Investigation
	መራመር	ተመራመረች, አመራመረ, ተመራመራችሁ, etc.	
	ምርምር	ምርምሩ, የምርምራቸውን, በምርምሯ, etc.	
	መርማር	መርማሪ, ከመርማሪው, መርማሪዋ, etc.	
	መርምር	እንመርምር, ይመርምሩ, ተመርምራ, etc.	
	መርመር	ይመርመር, ትመርመር, እንመርመር, ልመርመር, etc.	
	መረምር	እንመረምራለን, ሲመረምር, ይመረምራሉ, ስትመረምር, etc.	

morphological structure of words is known. For example, the stopwords ስለ /silə ‘about’/, ከ /kə ‘from’/, and አል /ʔəli ‘not’/ do not appear as a standalone word as shown in the following sample words. The word ስለመጣ /silamət’a ‘since he brought’/, ከልብ /kəlibi ‘from heart’/, አልመጣ /ʔəlimət’a ‘did not come’/ are equivalent to ስለ+አመጣ, ከ+ልብ, and አል+መጣ, respectively. As there could be several sequences of affixes representing various linguistic functions, words can appear in various morphological structures. As a result, Amharic stopwords usually have many variants. For example, the stopword ሌላ /lela ‘other’/ has variants የሌላ /jələla/, ሌሎች /lelotʃi/, በሌላኛው /bəlelanawi/, etc. This indicates that stopword identification and term representation in Amharic IR demands a different consideration than the conventional trend. It means that one could not work with the surface forms of words to identify and remove stopwords. Therefore, we removed them after applying morphological analysis on documents and queries using a stopword list. The stopword list itself is constructed from a corpus after applying morphological analysis. We removed stopwords from stem-based and root-based corpora using our stem-based and root-based stopword lists, respectively.

4.4 Indexing

In our system, document processing involves text preprocessing, morphological analysis, stopword removal and indexing. As a result of these processes, we obtain indexed documents. To test the impact of morphological analysis on Amharic IR, word-based, stem-based and root-based indexes were created using Lemur¹ toolkit. The stem-based index was created using basic stems of words while the root-based index was created using the root of words. The number of root-based index terms is less than or equal to stem-based index terms. However, the frequency of a root is greater than or equal to the frequency of the corresponding stem as root form conflates all variants of a word to a single common form. Accordingly, the frequency of terms accurately computed

¹ <http://www.lemurproject.org>.

using root forms, which means that index term selection could be appropriately made by making use of the root forms of words.

4.5 Word Embedding

One of the main objectives of this work is to create an efficient model on a large Amharic dataset and investigate the impact of query expansion using term embedding on Amharic IR retrieval effectiveness. To this effect, we propose four neural network models using word2vec: stem-based with Continuous Bag of Words (CBOW), stem-based with skip-gram algorithm, root-based with CBOW and root-based with skip-gram algorithm. Accordingly, four vector space models are generated, which are used for expanding stem-based and root-based query terms based on semantic similarity of words in stem-based and root-based corpora, respectively. The similarity sim between a query term q and a corpus word d is computed using cosine similarity as shown in Eq. (1).

$$sim(q, d) = \frac{\sum_i qi \cdot di}{\sqrt{\sum_i qi^2 \sum_i di^2}} \quad (1)$$

where qi is vector representation of the i^{th} query term and di is the vector representation of the i^{th} word in the corpus. The top 5 most related terms are used to expand query terms.

4.6 Query Expansion

The root-based morphological analysis addresses variation among word variants during exact matching between Amharic documents and queries. However, matching only keywords may not accurately reveal the semantic similarity between a query and a document. To resolve this issue and optimize Amharic IR system, we performed query expansion using vector space model and WordNet. Semantically related terms to each non-stopword of user query are identified based on the word vector space and WordNet. Since there is no publicly available Amharic WordNet, we build the resource to be used only for the title of the topics from 2AIRTC [24]. The WordNet is organized to include terms' synonyms, hypernyms, and hyponyms relationships. The stem-based and root-based morphological analyses are carried out on words included in the WordNet and a user query. For query expansion, the stem or root of semantically related words from the WordNet are added to the original set of query term(s).

4.7 Matching and Ranking

Query term vector for searching is constructed after a query is subjected to preprocessing, morphological analysis, and stopword removal. Here, we applied both semantic-based and exact vocabulary term matching. The system searches documents that contain query terms and semantically related words (i.e. expanded terms). Searching for relevant documents is carried out by matching query terms (representing information need of users) with index terms (representing documents). As documents and query terms are represented using stem and root forms of words, the stem-based query terms are matched

against stem-based index terms whereas root-based query terms are matched against root-based index terms. In IR, a given user information need does not uniquely identify one document in the corpus. Instead, many documents might match a query but with different degree of relevancy. For a given query Q and a collection of retrieved documents D , the Lemur toolkit ranks retrieval results based on their possible relevance. The document length and number of matching query terms are taken into consideration. OKAPI BM25 score ranks documents based on Eq. (2).

$$score(D, Q) = \sum_{i=1}^n IDF(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{avgdl}\right)} \quad (2)$$

where $f(q_i, D)$ is q_i '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. The variables k_1 and b are free parameters whereas $IDF(q_i)$ is the inverse document frequency weight of the query term q_i . For language modeling, the similarity between a document D and a query Q is measured by the Kullback-Leibler (KL) divergence between the document model $D\theta$ and the query model $Q\theta$. The KL divergence ranking function captures the term occurrence distributions and it is computed using Eq. (3) as:

$$KL(Q\theta, D\theta) = \sum_{w \in V} p(w|Q\theta) \log \frac{p(w|Q\theta)}{p(w|D\theta)} \quad (3)$$

where w is word, v is word vector, $p(w|Q\theta)$ is estimated query term, $p(w|D\theta)$ is the smoothed probability of a term seen in the document. The ranking of the results of the proposed IR system was evaluated by precision, recall, mean precision and normalized discounted cumulative gain (NDCG). *Precision* is used to measure how many of retrieved documents are relevant and it is computed as:

$$Precision = \frac{\text{relevant item retrieved}}{\text{retrieved items}} \quad (4)$$

Recall measures the ability of an IR system to retrieve all relevant items and it is computed as:

$$Recall = \frac{\text{relevant item retrieved}}{\text{relevant items}} \quad (5)$$

Mean Average Precision (MAP) indicates a single-figure measure of quality across multiple queries. The *MAP* value is obtained by taking the mean of the Average Precision $Pav(q_i)$ over all the queries in the set Q and it is computed as:

$$MAP = \frac{1}{|Q|} \sum_{q_i \in Q} Pav(q_i) \quad (6)$$

Normalized Discounted Cumulative Gain (*NDCG*) is used to measure the position of relevant documents in the retrieval set. It is calculated as:

$$NDCG = \frac{DCG}{IDCG} \quad (7)$$

where *DCG* is discount cumulative gain, *IDCG* is the ideal discounted cumulative gain and it is the maximum possible value. These measures are valued between 0 and 1.

5 Construction of Amharic IR Resources

5.1 Context-Based Morphologically Annotated Corpora

Segmenting a word into its morphemes and extracting its base is crucial in many applications. In this work, Amharic surface words are segmented into their morphemes by analyzing the internal structure of words and their contexts. The annotation is made semi-automatically using Amharic lexicons built by Yeshambel *et al.* [10]. For comparison of stem-based and root-based text representations, we created the stem-based and root-based corpora from the same document collection. Words are morphologically segmented into affixes and basic stems or roots. The general annotation structure for a word W is represented as:

$$[p_ * w[_s]^*$$

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 ስማቸው /*simatfəwi*/ can be annotated as follows.

$$\underbrace{\text{ስም}}_w \underbrace{\text{አቸው}}_s$$

A single word may have multiple annotations when annotated with a single base form. However, among multiple annotations, only one of them could be relevant in a given context. For example, the root-based annotation of the word ስማቸው /*simatfəwi*/ could be ስም አቸው /*simi_ʔətfəwi*/ ‘their name’/, ስማቸው /*simatfəwi*/ ‘a person having a name ‘Simachew’/, ስም አቸው /*s-m_ʔə_ʔətfəwi*/ ‘you listen them’/, and ስ-ፊ-ም አቸው /*s-f-m_ʔə_ʔətfəwi*/ ‘having that she kissed them’/. This may lead to incorrect retrieval results. Thus, we identify the context of the word in a sentence during the annotation process. We annotate each word in the corpus with a single annotation by taking the context into consideration and context-based morphologically annotated corpora is constructed for stem-based and root based text representation. Depending on morphological structures, Amharic words can be categorized as derived and non-derived from verbs. Words derived from verbs may have various word classes, but they are morphologically generated from verbal stems or roots. The root forms of such words are represented only by radicals. On the other hand, words that are non-derived from verbs have root forms that contain radicals and vowels. The process of context-based morphological annotation of different word classes is presented as follows.

Stem-Based Morphologically Annotated Corpus. The stem-based morphological annotation segments word forms into more general representation known as basic stem and affixes. The majority of Amharic words are composed of stems and attached affixes. IR systems use the stems of words during indexing and term matching, and thus, we segment the stems of words from the rest of morphemes.

Stem-Based Annotation of Words Derived from Verbs: The base of many verbs, nouns, adjectives, and adverbs are verbal roots. Stems can be generated from a single verbal root and many words can be generated from a single stem by attaching affixes. For example, the verb ስላልፈጠኑ /silalifət'ənu/ 'since they haven't been fast', the noun ከፍጥነታችን /kəfit'inətətfini/ 'our speed', the adjective እንደፈጣኖቹ /ʔinidəfət'anotfu/ 'like the fast ones' and the adverb በፍጥነት /bəfit'inəti/ 'quickly' are derived from the verbal root ፍ-ጥ-ን /f-t-n/. These words are generated from three stems (ፈጠን /fət'əni/, ፈጣን /fət'ani/ and ፍጥን /fit'ini/) and their stem-based annotation is shown below.

ስላልፈጠኑ ስለ_ክል_ፈጠን_ኩ silə_ʔəli_fət'əni_ʔu [pre]-[neg]-[stem]-[3,pl]	ከፍጥነታችን ከ_ፍጥን_ኧት_አችን kə_fit'ini_ʔəti_ʔətfini [pre]-[stem]-[nom]-[1,pl,pos]
እንደፈጣኖቹ እንደ_ፈጣን_አች_ኩ ʔinidə_fət'ani_ʔotfi_ʔu [pre]-[stem]-[pl]-[def]	በፍጥነት በ_ፍጥን_ኧት bə_fit'ini_ʔəti [pre]-[stem]-[nom]

Stem-Based Annotation of Words Not Derived from Verbs: Amharic words may also be generated from primary nouns, adjectives, adverbs and functional words. Such words are not formed from verbal roots, and their stem representation is different from that of verbal stems as variants of words that are not derived from verbal stems have a single common basic stem. For example, words derived from the primary noun ሀገር /hagəri/ 'country' include ለሀገራችን /ləhagəratʃini/ 'for our country', ስለሀገሪቱ /siləhagəritu/ 'about the country', ሀገራዊ /hagərawi/ 'national', etc. The stem-based annotations of these words are presented as follows.

ለሀገራችን ለ_ሀገር_አችን lə_hagəri_ʔətfini [pre]-[stem]-[1,pl,pos]	ስለሀገሪቱ ለ_ሀገር_ኢት_ኩ lə_hagəri_ʔəti_ʔu [pre]-[stem]-[f]-[def]	ሀገራዊ ሀገር_አዊ hagəri_ʔəwi [stem]-[adj]
---	---	---

Root-Based Morphologically Annotated Corpus. Root-based morphological annotation segments the roots of words from other affixes. The annotation helps to investigate the impact of root-based text representation on Amharic IR. The root annotation process for verbal words differs from that of others as presented below.

Root-Based Annotation of Words Derived from Verbs. Verbal words are words derived from verbal roots by inserting new character, palatalizing one or more characters, changing the shape of one or more characters, or adding affixes. For example, the verb ስላልፈጠኑ /silalifət'ənu/ 'since they haven't been fast', the noun ከፍጥነታችን /kəfit'inətətfini/ 'from

our speed’/, the adjective እንደፈጣኞቹ /ʔinidəfət’anotʃu/ ‘like the fast ones’/ and the adverb በፍጥነት /bəfɪt’inəti/ ‘quickly’/ are derived from the verbal root ፍ-ጥ-ን /f-t'-n/. The root-based annotations are shown below.

ስላልፈጠኑ ስለ_አል_ፍ-ጥ-ን_ኡ silə_ʔəli_f-t'-n_ʔu [pre]-[neg]-[root]-[3,pl]	ከፍጥነታችን ከ_ፍ-ጥ-ን_ኧት_አችን kə_f-t'-n_ʔəti_ʔətfɪni [pre]-[root]-[nom]-[1,pl,pos]
እንደፈጣኞቹ እንደ_ፍ-ጥ-ን_አች_ኡ ʔinidə_f-t'-n_ʔətfɪ_ʔu [pre]-[root]-[pl]-[def]	በፍጥነት በ_ፍ-ጥ-ን_ኧት bə_f-t'-n_ʔəti [pre]-[root]-[nom]

Root-Based Annotation of Words Not Derived from Verbs: The root and stem forms are the same for words that are generated from primary nouns, adjectives, adverbs and functional words. The root forms may contain vowel in addition to radicals. For example, the words ለሀገራችን /ləhagəratʃɪni/ ‘for our country’/, ስለሀገሪቱ /siləhagəritu/ ‘about the country’/, ሀገራዊ /hagərawi/ ‘national’/, etc. are derived from the primary noun ሀገር /hagəri/ ‘country’/. The root-based annotations of these words are presented as follows.

ለሀገራችን ለ_ሀገር_አችን lə_hagəri_ʔətfɪni [pre]-[root]-[1,pl,pos]	ስለሀገሪቱ ለ_ሀገር_ኡት_ኡ lə_hagəri_ʔəti_ʔu [pre]-[root]-[f]-[def]	ሀገራዊ ሀገር_አዊ hagəri_ʔəwi [root]-[adj]
--	--	--

5.2 Stopword List Construction

We remove Amharic stopwords from the vocabulary using a predefined list constructed based on stem and root forms. For the sake of comparison between stem-based and root-based text retrieval, both types of stopword lists were created from the annotated corpora based on morpheme statistics involving frequency, mean, variance, and entropy. The values of frequency, variance, entropy and mean of each morpheme in the corpus were used while constructing the stopword list. The top 250 morphemes based on the values of frequency, variance, entropy and mean are selected to create corpus-based stopword lists. However, the final stopword list also contains a few words which were selected manually from other sources considering the nature of the language. In total, 222 morphemes are included in each stopword list. The identified stopwords include prepositions (e.g. ወደ /wədə/ ‘to’/, ስለ /silə/ ‘about’/, እስከ /ʔiskə/ ‘up to’/, በ /bə/ ‘by’/, ከ /kə/ ‘from’/, etc.), conjunctions (e.g. እና /ʔina/ ‘and’/, ይሁን እንጅ /yihuni ʔinidzi/ ‘however’/, እዚህ /ʔizih/ ‘here’/, etc.), negation markers (አል...ም /ʔəli...mi/ ‘not’/), indefinite articles (አንድ /ʔənidi/ ‘an’/), auxiliary verbs (እል /ʔ-l/ ‘say’/, ን-በ-ር /n-b-r/ ‘was’/, etc.), ወዘተ /wəzətə/ ‘and so on’/, etc. The

stem-based stopword list may contain multiple stems for variants of a word while the root-based list contains only one root form for variants of a word.

6 Experiment

6.1 Implementation

We carried out different experiments on 2AIRTC collection [24]. Python was used to implement preprocessing tasks whereas Lemur toolkit was used for indexing and retrieval. The retrieval effectiveness was evaluated automatically using *trec_eval* tool which can compute many evaluation measures². LM and BM25 models were used as retrieval models.

6.2 Experimental Results

Retrieval with LM and BM25. LM is a popular model for the development of IR systems, but it has not been used in previous Amharic IR ones. Many of them are rather based on vector space model [13, 20, 32, 34]. Here, we investigated the effect of LM model on Amharic IR and compared it with BM25. As shown in Table 3, LM performs slightly better than BM25 model. This is potentially because of the capability of LM to capture term dependency and estimate the probability distribution of a query in each document. This means LM is more suitable retrieval model for Amharic language.

Table 3. Comparison of LM and BM25.

Model	Average precision	R-precision	NDCG	Bpref
BM25	0.67	0.64	0.83	0.64
LM	0.70	0.65	0.86	0.66

The Effect of Stopword Removal on Amharic IR. Experiments were conducted to investigate the effect of morpheme-based stopword removal on Amharic IR. The retrieval effectiveness of stem-based and root-based text representations with and without stopwords are shown in Table 4. As shown in the table, removing stopwords has a positive impact both in stem-based and root-based retrieval using LM.

Retrieval Without Query Expansion. The retrieval effectiveness of the proposed Amharic IR model without query expansion using LM is presented in Table 5. As shown in the table, root-based retrieval is better than stem-based retrieval as the root-based text

² http://trec.nist.gov/trec_eval.

Table 4. Stem-based and root-based retrieval with and without stopwords on 2AIRTC.

Metrics	Stem-based		Root-based	
	With stopword	Without stopword	With stopword	Without stopword
AMP	0.14	0.51	0.24	0.70
NDCG	0.37	0.71	0.50	0.86
Bpref	0.15	0.48	0.27	0.66
R-prec	0.17	0.49	0.29	0.65

representation maps all variants to a single common form and can reject non-relevant documents better than stem-based and word-based text representations. The word-based and stem-based methods miss more relevant documents since they cannot handle some morphological variations. The retrieval effectiveness of the three text representations decreases from precision @5 documents to precision @20 due to scarcity of relevant documents in the test collection.

Table 5. Retrieval effectiveness based on the three text representations.

Text representation	Precision					NDLG
	P@5	P@10	P@15	P@20	MAP	
Word	0.56	0.49	0.44	0.40	0.43	0.47
Stem	0.62	0.53	0.47	0.43	0.57	0.71
Root	0.79	0.70	0.61	0.55	0.70	0.86

The overall recall and precision values of stem-based and root-based text representations are shown in Fig. 2 which has been taken from our previous study [33]. The blue line depicts the root-based retrieval effectiveness whereas the red line represents the stem-based retrieval results without query expansion. It can be seen that the retrieval effectiveness of root-based representation outperforms stem-based one.

Semantic Retrieval Using Word Embedding. CBOW and skip-gram learning algorithms were trained by adjusting the parameter settings into the following same values: *vector size* (300), *min_count* (7), *iter* (400), *alpha* (0.05) and *negative* (20). By experiment, we found that the best performing window size of CBOW (resp. skip-gram) are 3 (resp.7). Recall and precision of semantic retrieval using word embedding technique on stem-based and root-based corpora are shown in Figs. 3 and 4. The green curve depicts stem-based (Fig. 3) and root-based retrieval (Fig. 4) without query expansion and the remaining are after query expansion using word embedding with CBOW and skip-gram algorithms. The retrieval effectiveness of stem-based query expansion based on both algorithms are almost similar. In the case of root-based representation, CBOW

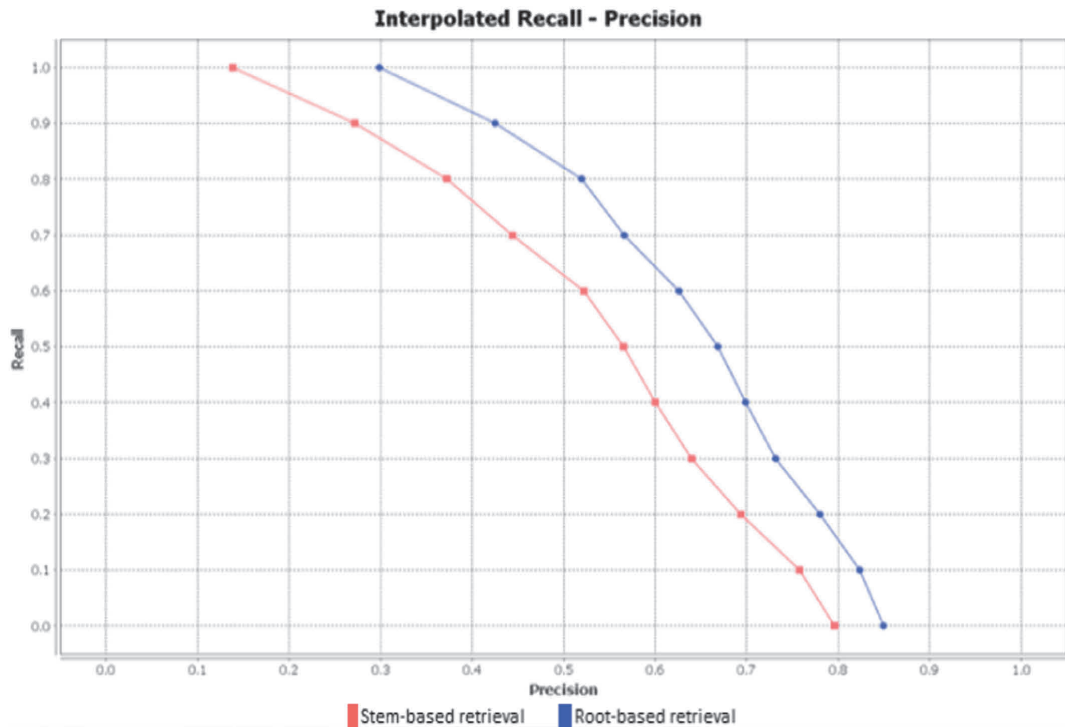


Fig. 2. Recall-precision curves of stem-based and root-based retrieval [33].

model slightly outperforms skip-gram method. However, the retrieval effectiveness after query expansion is reduced in stem-based and root-based representation. A statistical test is made for root-based query expansion using CBOW and Skip-gram models (see Table 6).

Table 6. Statistical test for root-based query expansion using CBOW and Skip-gram models.

Statistical test	Average precision	R-precision	NDCG	Bpref	P@5	P@10	P@20
t-test	0.7123	0.9564	0.7711	0.7778	0.6487	0.8413	0.9282
Randomized test	0.7043	0.9597	0.7643	0.7763	0.6647	0.8443	0.9345
Sign test	0.5270	0.9924	0.7718	0.9183	0.6847	0.7576	0.8356

Semantic Retrieval Using WordNet. The retrieval effectiveness of stem-based and root-based text representations are investigated with the application of query expansion using WordNet. Experimental results are shown in Figs. 5 and 6. The red curve is retrieval without query expansion and the others are semantic retrieval with query expansion using WordNet.

6.3 Discussion

Comparison of Root and Stem for Retrieval. Both stem-based and root-based text representations improve Amharic retrieval effectiveness in comparison to word-based

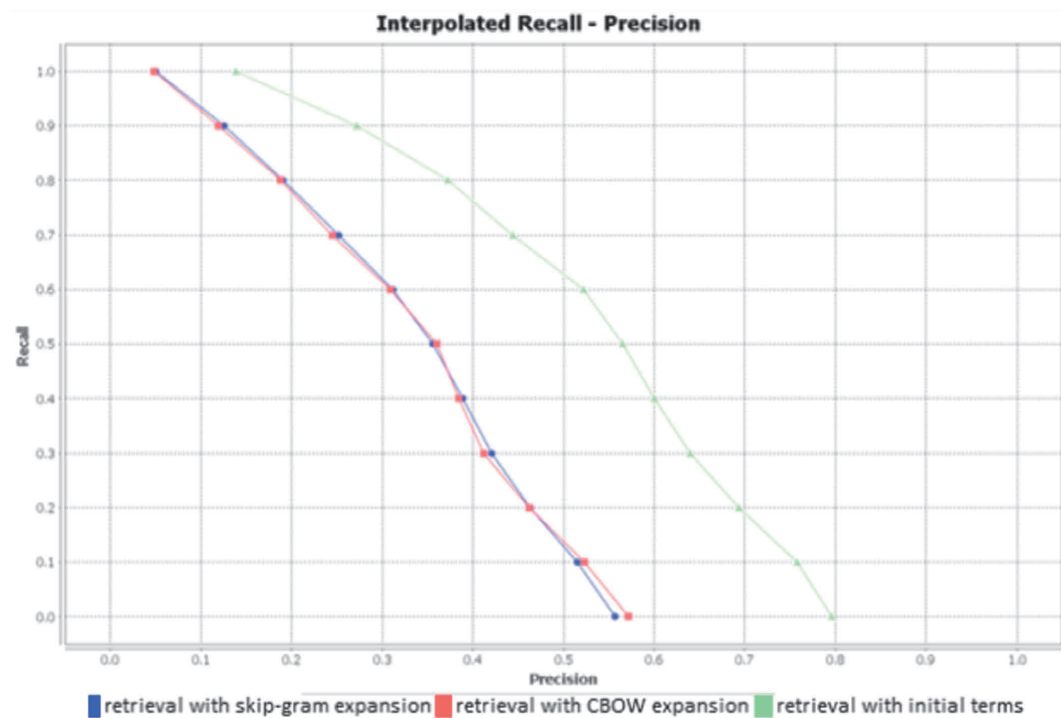


Fig. 3. Recall-precision curves of stem-based semantic retrieval using word embedding.

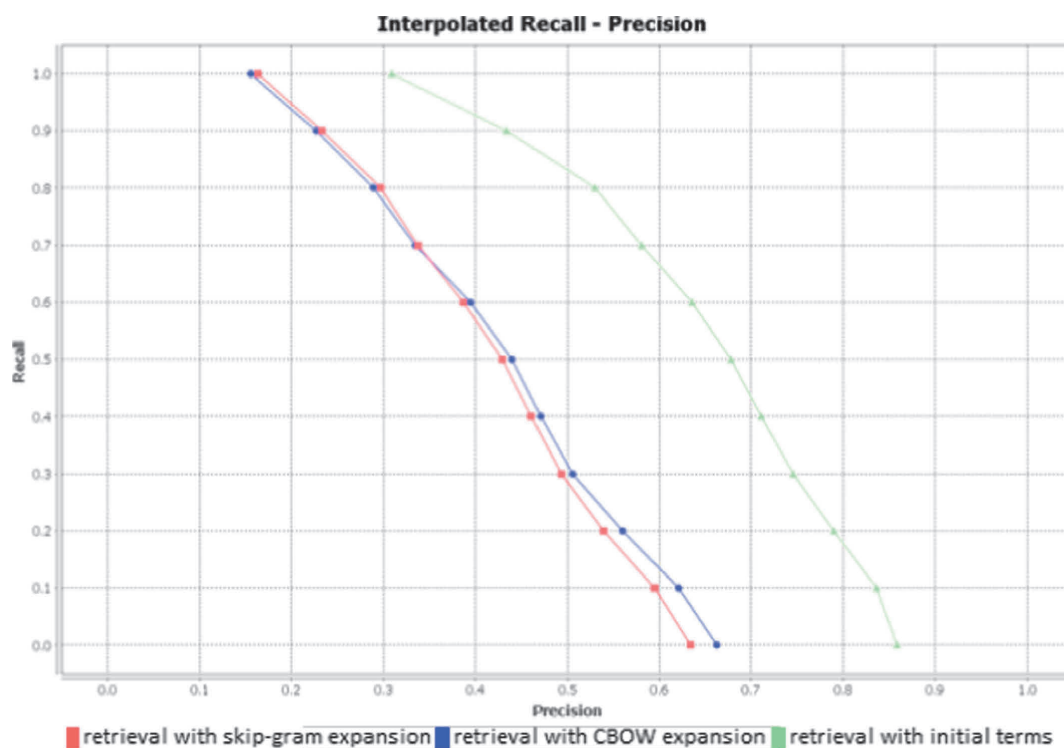


Fig. 4. Recall-precision curves of root-based semantic retrieval using word embedding.

text representation. Root-based retrieval is better than stem-based possibly due to the following three reasons.

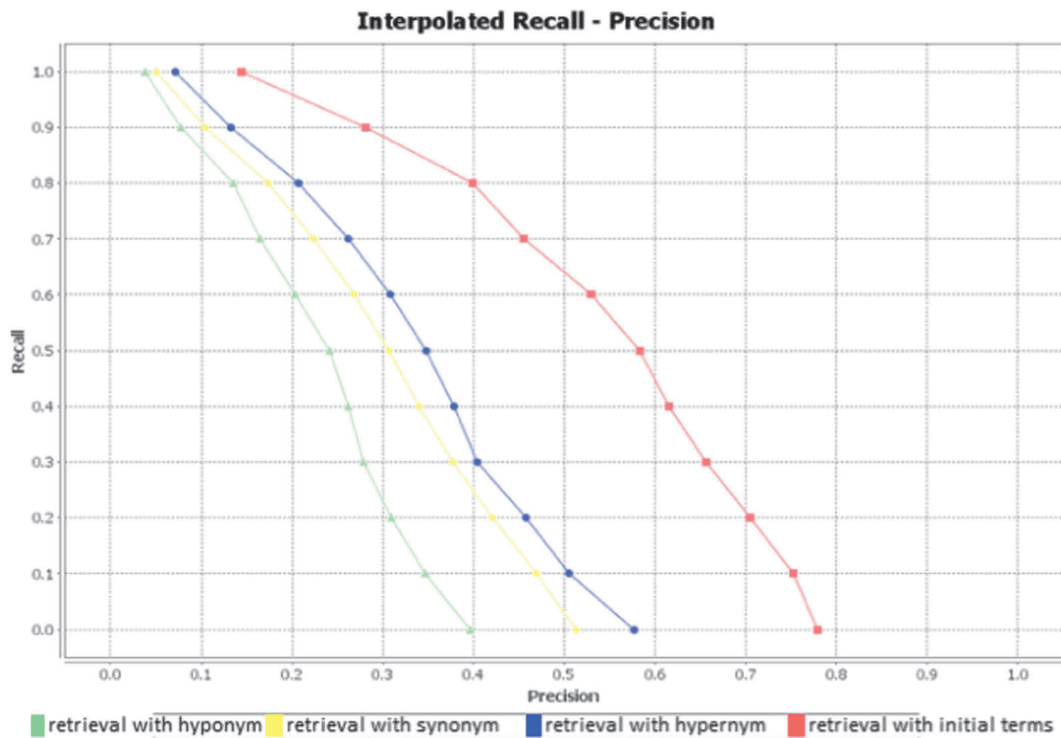


Fig. 5. Recall-precision curves of stem-based retrieval using WordNet.

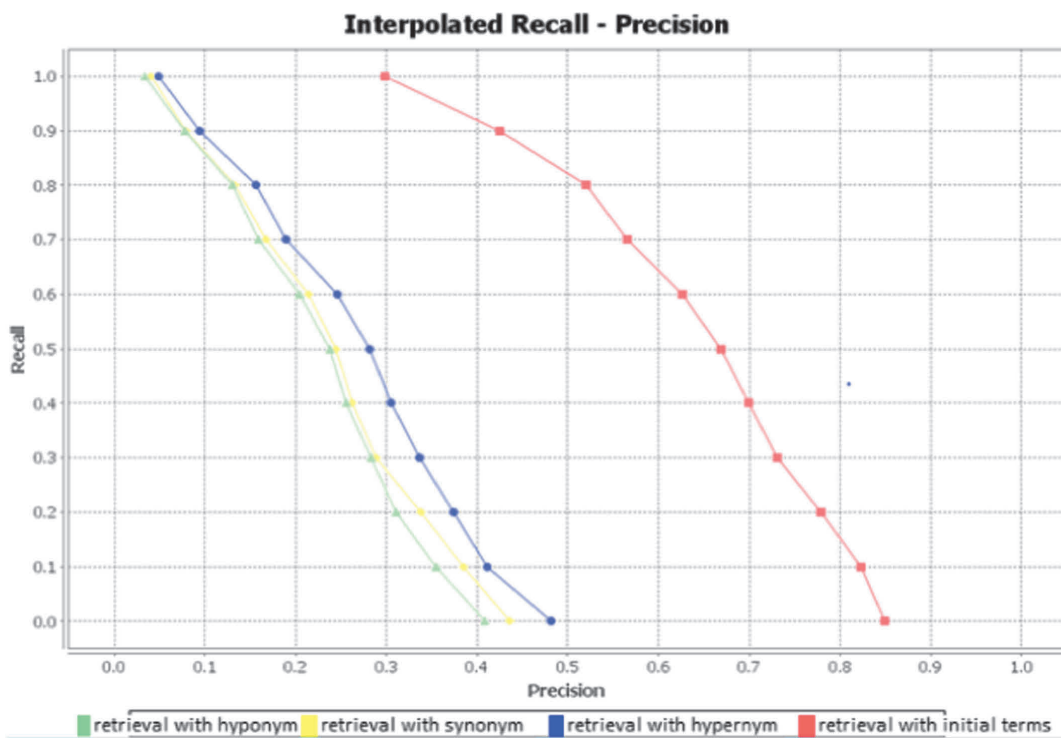


Fig. 6. Recall-precision curves of root-based retrieval using WordNet.

First, variants of a word have a single common root but might have more than one stem. For example, the stems of morphological variants ጠንካራ /*t'anakara*/, ጥንካሬ /*t'inikare*/, ጠንካራ /*t'anikara*/, እንጠንካር /*ʔinit'anikiri*/ and አጠናካረ /*ʔat'anakara*/ are

ጠነከር /t'anəkəri/, ጥንካር /t'inikari/, ጠንካር /t'anikari/, ጠንክር /t'anikiri/ and ጠናከር /t'anakəri/, respectively. This creates term mismatch with each other. As a result, stem-based text representation is unable to retrieve all relevant documents and compute the actual term frequency which results in loss of the rank of retrieved relevant documents. However, all variants have one common root ጥ-ን-ከ-ር /t'-n-k-r' 'strong'/. Therefore, the root-based representation can return more relevant documents than stem-based representation and can compute and increase the actual term frequency which usually leads to better retrieval result at correct rank.

Second, root forms do not conflate semantically unrelated words to a common form. However, the stem-based text representation sometimes conflates semantically unrelated words. For example, ቀን /k'anɪ/ is the stem of the verb ቀና /k'ana' 'upright' or 'become jealous' / and the noun ቀናት /k'anati' 'days'/. However, their roots are ቅ-ን /k'-n/ and ቀን /k'anɪ/, respectively. Many cases like these occur in the language, and stem-based text representation increases word ambiguities than root-based text representation. Thus, the stem-based text representation is unable to filter out some non-relevant documents.

Third, stem-based retrieval depends largely on user query formulation. Different users will certainly construct the same information need using different word variants. For example, the query 'deforestation' can be constructed as የደን መጨፍጨፍ /jədəni mətʃəfɪtʃəfi/ or የደን ጭፍጨፋ /jədəni tʃifitʃəfa/. After the stem-based morphological analysis, the two queries have stem terms ደን መፍጨፍ and ደን ጭፍጨፍ, respectively. As a result of variation of the second term, the system will return different results in different ranks. Therefore, stem-based text representation performs differently in our test collection. However, the root-based representation performs equally for all the variants of the query terms as the two queries have the same root terms ደን ጭ-ፍ-ጭ-ፍ.

Comparison with Other Amharic IR Systems. Few Amharic IR systems have been developed so far. Some of them are based on stems [13, 20, 34]; while some others are based on citation forms [32]. The effects of stem-based and root-based text representations are investigated on Amharic IR [12]. They found that the stem-based representation is better than the root-based representation. They stated that root-based representation maps semantically unrelated Amharic words. However, roots were represented incorrectly in their research. For example, the word ጥጥ /t'it'i' 'cotton' / and ጠጣ /t'ət'a' 'drink' / were represented incorrectly as ጥጥ even though their correct roots are ጥጥ /t'it'i/ and ጥ-ጥ /t'-t'/, respectively. Furthermore, the roots of some verbal stems are represented incorrectly. For example, መታ /məta' 'hit' / and ሞተ /motə' 'die' / were mapped incorrectly to a common root ሞት even though the correct roots are ሞ-ት and ሞ-ው-ት, respectively. On the other hand, they remove vowels from all types of words leading to conflation many semantically unrelated words to the same form. The Google Amharic search engine retrieves different documents in different ranks for basic stems and their derived stems though they are morphological variants. For example, Google search results of the queries ከብራት /sibirati' 'being broken' / and መሰበር /məsəbəri' 'the process of being broken' / are different though the same concept is expressed via these two variants. In our work, on the contrary, the stem-based text representation considers only basic stems and provides the same retrieval results for both basic stems and derived stems. We use root-based text representation as it conflates all variants of words to a single common form. In

summary, previous studies that recommended the use of stems made their conclusions without thorough investigation on the applicability of roots. Many of them suggested stem-based as the best option. However, due to the complexity of the language stem-based representation does not work well. In this work, we have shown that the roots are better than stems for Amharic IR. This is a new finding which was not looked at in previous work.

Comparison of Conventional and Semantic Amharic IR. Even though the proposed morphological analysis (i.e. root-based) has positive impact on Amharic IR, query expansion using word embedding and WordNet does not improve the performance of the system due to term ambiguity. Ambiguous terms are prevalent in the language due to its complex morphology. For example, the term አለሙ /*ʔələmu*/ can mean ‘the world’, ‘they targeted’, ‘they dreamed’, ‘they developed’ or a person with the name ‘Alemu’. Accordingly, the term አለሙ /*ʔələmu*/ needs to be expanded based on the context of the term in a given query. This affects the retrieval effectiveness of the proposed Amharic IR system. Since Amharic word sense disambiguator is not available yet, we did not integrate disambiguator into our proposed semantic retrieval system. Moreover, stem-based representation may not expand query terms though its semantically related words are in the WordNet. This is because a word can have several stems but there could be only one form representing the word in the WordNet. In this case, the plausible way to organize Amharic WordNet is to make use of root form as it can represent all variants of a word by a single common form. This could not be achieved by stem-based representation. Thus, root-based expansion could work well if word sense disambiguator was integrated in this work. Moreover, the case of word embedding, variants of a given word in the corpus might co-occur with variants of a given semantically related word in different forms. For example, the word ደን /*dəni*/ might co-occur with ጭፍጨፋ /*tʃʻifitʃʻəfa*/, መጨፍጨፍ /*mətʃʻəfitʃʻəfi*/, ጨፍጫፊ /*tʃʻəfitʃʻəfi* /, and ጨፍጭፍ /*tʃʻəfitʃʻifi* / within a specified window size. As these words have different stems, the actual co-occurrence frequency based on stem could not be computed correctly. As a result, the similarity between a query term and its semantically related word is lower which affects stem-based semantic retrieval. Furthermore, some expanded terms are related to a query term syntactically rather than semantically. For instance, the expanded terms for the proper noun በቀለ /*bəkʻələ*/ are ጥላሁን /*tʻilahun*/, አበራ /*ʔəbərə*/, መስፍን /*məsifin*/, ማሞ /*mamo*/ and ንጉሴ /*niguse*/ where their meanings are completely different. Consequently, retrieval using expanded terms sometimes returns more non-relevant documents than the original query retrieval. The overall retrieval effectiveness using expanded terms is lower than retrieval with only original query terms. The other possible reason for lower performance could be the small size of the corpus. However, a promising result was reported in previous Amharic IR research even using stems [20].

Comparison of Our Stopword List with Others. Few researches were conducted to build Amharic stopwords. However, classical methods that have been used in many morphologically simple languages such as English are applied without considering the characteristics of Amharic. For example, stopword lists constructed by Mindaye *et al.* [13] and Samuel and Bjorn [25] contain variants of a word. However, it is challenging to list all the variants of stopwords. Alemayehu and Peter [26] created stopword list based

on stem. Though stems are better than word forms to construct Amharic stopwords, it is not the plausible way because of the existence of multiple stems for variants of a stopword. In our case, all the variants of stopword have a single common form. For example, the stopword list created by Alemayehu and Peter [26] would contain two stems (ነባር /*nəbəri*/, ነባር /*nəbari*/) which are variants of a single word. However, in our case all variants of the stopword are represented by single root ን-ብ-ር /*n-b-r* ‘was’/.

7 Conclusion

Amharic has complex morphology which poses tremendous challenges for NLP and IR. In this work, we evaluate the existing Amharic NLP tools and resources, and investigate the implications of the morphological complexity on Amharic IR. After analyzing the gaps, we constructed standard resources and proposed a new Amharic IR system that takes the morphology of the language into consideration. The resources that we constructed are Amharic stopword list and context-based morphologically annotated corpora. They are made publicly accessible to the research community. Furthermore, stem-based and root-based morphological features were considered to construct resource, corpora, and develop Amharic IR system. Our findings indicate that root is the optimal form of word representation for Amharic IR development and resource construction. We also investigated semantic-based query expansion based on word embedding and WordNet. We exploited the deep learning models (i.e. CBOW and Skip-gram) and WordNet to deal with term mismatch in Amharic IR though negative results were obtained due to prevalent term ambiguity. Further research on Amharic IR needs to be conducted by integrating Amharic word sense disambiguation so that only relevant terms are considered during query expansion of words having multiple interpretations.

References

1. Xu, J., Croft, W.: Query expansion using local and global document analysis. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 4–11. ACM (1996)
2. Ben, W., Karaa, A.: A new stemmer to improve information retrieval. *Int. J. Netw. Secur. Appl. (IJNSA)* **5**(4), 143–154 (2013)
3. Coustié, O., Mothe, J., Teste, O., Baril, X.: Meting: a robust log parser based on frequent n-gram mining. In: 2020 IEEE International Conference on Web Services (ICWS), pp. 84–88 (2020)
4. Jabbar, A., Iqbal, S., Tamimy, M.I., Hussain, S., Akhunzada, A.: Empirical evaluation and study of text stemming algorithms. *Artif. Intell. Rev.* **53**(8), 5559–5588 (2020). <https://doi.org/10.1007/s10462-020-09828-3>
5. Lavrenko, V., Croft, W.: Relevance based language models. In: SIGIR 2001, New Orleans, Louisiana, USA, pp. 260–267 (2001)
6. Xu, Y., Jones, G.J., Wang, B.: Query dependent pseudo-relevance feedback based on Wikipedia. In: SIGIR 2009, Boston, MA, USA, pp. 59–66 (2009)
7. Harb, H., Fouad, K., Nagdy, N.: Semantic retrieval approach for web documents. *Int. J. Adv. Comput. Sci. Appl.* **2**(9) (2011)

8. El-Mahdaouy, A., Ouatik, S., Gaussier, E.: Semantically enhanced term frequency based on word embedding for Arabic information retrieval. In: 4th IEEE International Colloquium Information Science and Technology (CiSt), pp. 385–389 (2016)
9. Abate, M., Assabie, Y.: Development of Amharic morphological analyzer using memory-based learning. In: Przepiórkowski, A., Ogrodniczuk, M. (eds.) NLP 2014. LNCS (LNAI), vol. 8686, pp. 1–13. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10888-9_1
10. Yeshambel, T., Mothe, J., Assabie, Y.: Morphologically annotated Amharic text corpora. In: Proceedings of 44th ACM SIGIR Conference on Research and Development in Information Retrieval, Online Conference, Canada, pp. 2349–2355 (2021)
11. Countrymeters: Ethiopian population (2021). <https://countrymeters.info/en/Ethiopia>. Accessed 02 Aug 2021
12. Alemayehu, N., Willett, P.: The effectiveness of stemming for information retrieval in Amharic. *Program Electron. Libr. Inf. Syst.* **37**(4), 254–259 (2003)
13. Mindaye, T., Redewan, H., Atnafu, S.: Design and implementation of Amharic search engine. In: Proceedings of the 5th International Conference on Signal Image Technology and Internet Based Systems, pp. 318–325 (2010)
14. Al-Hadid, Afaneh, S., Al-Tarawneh, H., Al-Malahmeh, H.: Arabic information retrieval system using the neural network model. *Int. J. Adv. Res. Comput. Commun. Eng.* **3**(12), 8664–8668 (2014)
15. Musaid, S.: Arabic information retrieval system-based on morphological analysis (AIRSMA): a comparative study of word, stem, root and morpho-semantic methods. Ph.D. dissertation, Computer and Information Science, De Montfort University, United Kingdom (2000)
16. Moukdad, H.: A comparison of root and stemming techniques for the retrieval of Arabic documents. Ph.D. dissertation, Graduate School of Library and Information Studies, McGill University, Montreal (2002)
17. Larkey, L.S., Ballesteros, L., Connell, M.E.: Light stemming for Arabic information retrieval. In: Soufi, A., Bosch, A.V., Neumann, G. (eds.) *Arabic Computational Morphology*, pp. 221–243. Springer, Dordrecht (2007). https://doi.org/10.1007/978-1-4020-6046-5_12
18. Ali, A., Mosa, E., Abdullah, B.: An intelligent use of stemmer and morphology analysis for Arabic information retrieval. *Egypt. Inform. J.* **21**(4), 209–217 (2020). <https://doi.org/10.1016/j.eij.2020.02.004>
19. Ornan, U.: A morphological, syntactic and semantic search engine for Hebrew texts. In: Proceedings of the ACL-2002 Workshop on Computational Approaches to Semitic Languages, Philadelphia, Pennsylvania, USA, pp. 1–10 (2002)
20. Getnet, B., Assabie, Y.: Amharic information retrieval based on query expansion using semantic vocabulary. In: Delele, M.A., Bitew, M.A., Beyene, A.A., Fanta, S.W., Ali, A.N. (eds.) *ICAST 2020. LNICSSITE*, vol. 384, pp. 407–416. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-80621-7_29
21. Fang, H.: A re-examination of query expansion using lexical resources. In: Proceedings of ACL-2008: HLT, Columbus, Ohio, USA, pp. 139–147 (2008)
22. Bagherid, E., Ensane, F., Al-Obeidat, F.: Neural word and entity embeddings for Ad hoc retrieval. *J. Inf. Process. Manag.* **54**, 657–673 (2018)
23. Demeke, G., Getachew, M.: Manual annotation of Amharic news items with part-of-speech tags and its challenges. *ELRC Working Papers* **2**(1), 1–16 (2006)
24. 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 (2020). https://doi.org/10.1007/978-3-030-58219-7_5
25. Samuel, E., Bjorn, G.: Classifying Amharic news text using self-organizing maps. In: Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, Michigan, USA, pp. 71–78 (2005)

26. Alemayehu, N., Willett, P.: Stemming of Amharic words for information retrieval. *J. Lit. Linguistic Comput.* **17**(1), 1–17 (2002)
27. Alemu, A., Asker, L.: An Amharic stemmer: reducing words to their citation forms. In: *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, Prague, Czech Republic, pp. 104–110. Association for Computational Linguistics (2007)
28. Sisay, F., Haller, J.: Application of corpus-based techniques to Amharic texts. In: *Proceedings of MT Summit IX Workshop on Machine Translation for Semitic Languages* (2003)
29. Amsalu, S., Gibbon, D.: Finite state morphology of Amharic. In: *5th Recent Advances in Natural Language Processing*, pp. 47–51 (2006)
30. Gasser, M.: HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya. In: *Conference on Human Language Technology for Development*, Alexandria, Egypt, pp. 94–99 (2011)
31. Mulugeta, W., Gasser, M.: Learning morphological rules for Amharic verbs using inductive logic programming. In: *Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL8/AfLaT2012)*, Istanbul, Turkey, pp. 7–12 (2012)
32. Argaw, A.A., Asker, L.: Amharic-English information retrieval. In: Peters, C., et al. (eds.) *CLEF 2006*. LNCS, vol. 4730, pp. 43–50. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74999-8_5
33. Yeshambel, T., Mothe, J., Assabie, Y.: Amharic document representation for adhoc retrieval. In: *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR*, pp. 124–134 (2020). <https://doi.org/10.5220/0010177301240134>. ISBN 978-989-758-474-9; ISSN 2184-3228
34. Munye, M., Atnafu, S.: Amharic-English bilingual Web search engine. In: *Proceedings of the 4th ACM International Conference on Management of Emergent Digital EcoSystems (MEDES 2012)*, Addis Ababa, Ethiopia, pp. 32–39 (2012)