Amharic Adhoc Information Retrieval System Based on Morphological Features

Tilahun Yeshambel, Josiane Mothe, Yaregal Assabie

applied sciences

MDPI

Article

Amharic *Adhoc* Information Retrieval System Based on Morphological Features

Tilahun Yeshambel [1,*], Josiane Mothe [2] and Yaregal Assabie [3]

[1] ITPhD Program, Addis Ababa University, Addis Ababa P.O. Box 1176, Ethiopia
[2] Université Jean-Jaurès, Université de Toulouse, Componsante INSPE, IRIT, UMR5505 CNRS, 118 Rte de Narbonne, F31400 Toulouse, France; Josiane.Mothe@irit.fr
[3] Department of Computer Science, Addis Ababa University, Addis Ababa P.O. Box 1176, Ethiopia; yaregal.assabie@aau.edu.et
* Correspondence: tilahun.yeshambel@uog.edu.et; Tel.: +251-912-990195

Abstract: Information retrieval (IR) is one of the most important research and development areas due to the explosion of digital data and the need of accessing relevant information from huge corpora. Although IR systems function well for technologically advanced languages such as English, this is not the case for morphologically complex, under-resourced and less-studied languages such as Amharic. Amharic is a Semitic language characterized by a complex morphology where thousands of words are generated from a single root form through inflection and derivation. This has made the development of Amharic natural language processing (NLP) tools a challenging task. Amharic *adhoc* retrieval also faces challenges due to scarcity of linguistic resources, tools and standard evaluation corpora. In this research work, we investigate the impact of morphological features on the representation of Amharic documents and queries for *adhoc* retrieval. We also analyze the effects of stem-based and root-based text representation, and proposed new Amharic IR system architecture. Moreover, we present the resources and corpora we constructed for evaluation of Amharic IR systems and other NLP tools. We conduct various experiments with a TREC-like approach for Amharic IR test collection using a standard evaluation framework and measures. Our findings show that root-based text representation outperforms the conventional stem-based representation on Amharic IR.

Keywords: information retrieval; *adhoc* retrieval; Amharic; complex morphology; corpus; resources

Citation: Yeshambel, T.; Mothe, J.; Assabie, Y. Amharic *Adhoc* Information Retrieval System Based on Morphological Features. *Appl. Sci.* **2022**, *12*, 1294. https://doi.org/10.3390/app12031294

Academic Editor: Valentino Santucci

Received: 9 November 2021
Accepted: 23 December 2021
Published: 26 January 2022

1. Introduction

Searching digital information on the Web or a large corpus has become part of the human daily life. Information Retrieval (IR) is concerned with searching relevant documents to a user's query from a document collection. Both the research community and the industry have been very active in this field for more than 60 years [1]. 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.

The ability of an IR system to retrieve relevant documents effectively should be tested and its performance should be evaluated systematically. Performance evaluation of an IR system is indeed very crucial for scientific progress [2,3]. IR has an old history of evaluation. In *adhoc* retrieval, where the task for the system is to retrieve relevant documents for a given query within a fixed size corpus, both the evaluation framework and effectiveness measures are well established. Test collections are the most widely used resources for performance evaluation [4]. A standard *adhoc* retrieval test collection consists of three components: a corpus of documents to be searched in, a set of users' information needs or topics, and the associated relevance judgments indicating which documents are relevant for which topics. A large number of shared tasks rely on such collections. Some of the well-known text collections and evaluation programs are Cranfield project [5], Text REtrieval Conference (TREC) and more specifically TREC *adhoc* [6], Cross-Language Evaluation Forum (CLEF) [3],

and NACSIS Test Collection for Information Retrieval (NTCIR) [7]. The IR international conferences such as TREC, CLEF, NTCIR, INEX, and FIRE are held based on their own test collections.

IR systems work based on documents produced in natural language and consequently, the performance of IR is affected by the characteristics of the language [8]. As a result, NLP has been employed to improve the effectiveness of IR systems. For example, NLP tools and resources provide a means to find index terms and query terms that improve search results. Linguistic variation has significant impact on IR effectiveness as it leads to the omission of relevant documents to users' queries [8]. This calls for the need of dealing with language specific issues to improve the performance of IR systems. Thus, NLP has long attracted the attention of the IR community [9–11]. The morphology, orthography, tokenization, syntax, semantics, and compound splitting of a language are some of the issues to be considered while developing an IR system. Many languages have several word 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 [12]. Performing simple matching between words generated from the same root is not sufficient while retrieving relevant documents to a user query [13]. Numerous researches have been conducted to develop IR system, to construct resources and corpora for resourceful languages such as English, French, and German. This is not, however, the case for under-resourced languages such as Amharic [14].

Amharic is one of the under-resourced languages where, in comparison with technologically advanced languages, few attempts have been made to develop Amharic NLP tools, applications, resources and corpora [15–17]. Among the few efforts are the development of stemmer [18,19], morphological analyzer [16,20–22], part-of-speech tagger [15,23,24], resources [25–27], and corpus [28–30]. However, the existing tools, resources and corpora are not fully functional, limiting their usability for IR [31]. Furthermore, the number of studies reported on Amharic IR and related applications are considerably limited compared to what is carried out for other languages such as English. The morphological complexity of Amharic creates a challenge on the development and retrieval effectiveness of Amharic IR. Thus, in order to come up with an effective IR system, one has to deal with the complex morphological features of the language which itself leads to complex grammatical structure. As a result, finding effective term (word) representation for documents and queries has been an issue of theoretical discussion in Amharic IR.

In many languages, the use of surface forms of words to represent documents and queries is not a choice due to the proliferation of words that can be generated from a single root form. Thus, the word forms that are considered for document and query representation are stems and roots. Although some efforts have been made to develop Amharic IR systems using stems, the effectiveness of the systems with respect to the use of various forms has not been systematically analyzed [32]. This current research analyzes the use of stems and roots for content representation and investigates their effects on Amharic IR.

It is noted that a reference collection, annotated corpora and scientifically built resources would assist to carry out more research and development works on Amharic IR and NLP. To advance research on the development of Amharic NLP tools and IR system, there should be scientifically built resources and test collections. The development of Amharic NLP tools is a non-trivial task because of its complex morphology. Thus, the creation of morphologically annotated Amharic corpus is a long-sought resource for the research community. Morphological annotation of a language is the process of assigning linguistic information such as part-of-speech (POS) and morphological features to recognize each word in a document [33]. Morphological analyzers can help to automatically annotate text corpus. With regard to Amharic, there have been few attempts to develop morphological analyzers [16,21,22]. These tools are at a prototype stage and they are also limited in scope. The unavailability of full-fledged low-level Amharic NLP tools makes it difficult to create annotated corpus automatically. Moreover, there is lack of standard reference collection making it a major impediment to the development of Amharic IR and NLP tools. The

objective of this research work is: (1) to promote publicly accessible Amharic resources including a standard test collection for IR, morphologically annotated corpora and compiled stopwords list; and (2) to design Amharic IR system by investigating the optimal representation of documents and queries considering both root-based and stem-based text representations.

The rest of this paper is organized as follows. Section 2 describes Amharic language and its morphology. Section 3 discusses related work. In Section 4, we present the proposed Amharic IR system architecture along with various options to represent Amharic documents and queries. Section 5 presents the evaluation framework of the proposed Amharic *adhoc* retrieval system including the presentation of the resources and corpora, experimental setup, and experimental results. In Section 6, we present our conclusion and put forward further research directions.

2. Amharic Language

2.1. Amharic Writing System

Amharic is the working language of the government of Ethiopia currently having an estimated population of over 117 million [34]. It is the second most commonly spoken Semitic language in the world after Arabic [35]. Although many languages are spoken in Ethiopia, Amharic is the lingua franca and the most literary language serving as a medium of instruction in the education system of the country for a long period. Amharic uses Ethiopic alphabet and has 34 base characters along with modifications on the respective base characters. The alphabet is conveniently written in a tabular format of seven columns. The first column represents the basic form with vowel አ /ə/ and the other six orders represent modifications with vowels in the order of ኡ /ʔu/, ኢ /ʔi/, ኣ /ʔa/, ኤ /ʔe/, እ /ʔɨ/, and ኦ /ʔo/. For example, the base character ቦ /bə/ has the following modifications: ቡ /bu/, ቢ /bi/, ባ /ba/, ቤ /be/, ብ /bɨ/, and ቦ /bo/. Furthermore, there are labialized characters such as ቧ /bwa/, ሟ /mwa/, ቷ /twa/, ቋ /qwa/, ሯ /rwa/, etc. The language has also its own punctuations such as '።' (full stop), '፣' (comma), '፤' (semicolon), '፥' (colon), '፦' (preface colon), etc. It also borrows some punctuation marks such as '?' and '!' from Latin alphabet. Although Ethiopic alphabet has numeral system, the Arabic numeral system is rather most commonly used by Amharic.

2.2. Amharic Morphology

Amharic has complex morphology where many words could be generated through inflectional and derivational processes from other word forms. Many surface forms of words can be formed from a base form through complex affixation, reduplication, and Semitic stem interdigitation. Amharic words can be marked for a number of functions. For example, a verb can be marked for a combination of person, case, gender, number, tense, aspect, and mood. An Amharic word might take up to four prefixes and five suffixes [18]. For example, the word ያላስተሳሰርኳቸውንና /jalasɨtəsasərɨkwatʃəwinɨna/ is constructed from four prefixes (የ /jə preposition/, አል /ʔəli negation/, አስ /ʔəsɨ causative/, ተ /tə reciprocal/); the perfective verb (አሰር /ʔəsərɨ 'tie'/); and four suffixes (ኩ /ku first person singular subject marker for perfect verb/, አቸው /ʔətʃəwɨ third person plural object marker for perfective verb/, ን /nɨ accusative/, and ና /na conjunction/). Thousands of words can be generated from an Amharic root or its stems by changing the shape of characters in a stem or root, inserting vowels between root radicals, reduplicating one of the characters of the stem or word itself, and by adding affixes on stems [22]. There are multiple stem templates for verbal words. However, all variants have a single root representation template. The majority of words in the language are generated from verbal roots. Nouns, adjectives, and adverbs can be derived from verbal roots. The derivation of words from verbal roots usually involves stem formation followed by word formation. Amharic roots are the base of stems which in turn are the base of many surface forms of words (see Table 1).

Table 1. Stem and word formation.

Root	Stem	Word in Surface Form
ን-ግ-ር /*n-g-r*/	ንገር /*nigəri*/ ናገር /*nagəri*/ ነገር /*nəgiri*/	ንገሩ /*nigəru*/, ንገረኝ /*nigərəɲi*/, መንገሬ /*mənigəre*/, etc. ተናገር /*tənagəri*/, መናገራችን /*mənagəratʃini*/, etc. ሊነግረኝ /*linəgirəɲi*/, ሲነግራቸው /*sinəgiratʃəwi*/, etc.
ል-ም-ድ /*l-m-d*/	ለመድ /*ləmədi*/ ላመድ /*lamədi*/ ልምድ /*limidi*/	ለመደ /*ləmədə*/, ተለመደ /*tələmədə*/, etc. መላመዱ /*məlaməduu*/, አላመዳቸ /*ʔəlamədətʃi*/, etc. ልምዶችን /*limidatʃini*/, በልምዱ /*bəlimidwa*/, etc.
ስ-ክ-ር /*s-k-r*/	ስካር /*sikari*/ ሰከር /*səkəri*/ ስከር /*sikəri*/	ስካሩ /*sikaru*/, በስካራቸው /*bəsikaratʃəwi*/, etc. ሰከረ /*səkərə*/, አሰከረ /*ʔəsəkərə*/, ሰከሩ /*səkəru*/, etc. መስከር /*məsikər*/, አለመስከሬ /*ʔələməsikəre*/, ስከሩ /*sikəru*/, etc.

Amharic verbal stems are formed from verbal roots by inserting vowels between radicals (consonants). For example, the verbal stems ንገር /*nigəri*/, ላመድ /*laməd*i/ and ሰከር /*səkəri*/ are derived from the verbal roots ን-ግ-ር /*n-g-r*/, ል-ም-ድ /*l-m-d*/ and ስ-ክ-ር /*s-k-r*/, respectively. The patterns of these stem formations from the roots are ን-እ-ግ-ኧ-ር /*n-i-g-ə-r-i*/, ል-አ-ም-ኧ-ድ /*l-a-m-ə-d-i*/, and ስ-ኧ-ክ-ኧ-ር /*s-ə-k-ə-r-i*/, respectively. Here ን /*n*/, ግ /*g*/, ር /*r*/, ል /*l*/, ም /*m*/, ድ /*d*/, ስ /s/, ክ /*k*/ and ር /*r*/ are consonants (C) whereas እ, ኧ, and አ are vowels (V). Depending on the types of verbs, stems are produced based on template structures in various forms such as perfective, imperfective, jussive, imperative, etc. [36]. For example, the template structures of tri-radicals verb types is presented in Table 2.

Table 2. Templates of tri-radical verbs types.

Verb Form	ስ-ብ-ር /*s-b-r*/		ፍ-ል-ግ /*f-l-g*/		ም-ር-ክ /*m-r-k*/	
	Stem	**Template**	**Stem**	**Template**	**Stem**	**Template**
Perfect	ሰበረ /*səbərə*/	CVCVC	ፈለገ /*fələgə*/	CVCVC	መረከ /*marəkə*/	CVCVC
Imperfect	ሰብር /*səbir*i/	CVCC	ፍለግ /*fəligi*/	CVCC	መርክ /*marik*i/	CVCC
Jussive	ስበር /*sibəri*/	CCVC	ፍለግ /*fəligi*/	CVCC	መርክ /*marik*i/	CVCC
Gerund	ሰብር /*səbir*i/	CVCC	ፍለግ /*fəligi*/	CVCC	መርክ /*marik*i/	CVCC
Infinitive	ስበር /*sibəri*/	CCVC	ፍለገ /*fiələgi*/	CVCVC	መረከ /*marəki*/	CVCVC

Word formation from verbal stems is completed by attaching different affixes such as infinitive, gerund, person, gender, number, case, passive, tense/aspect, genitive, accusative, mood, etc. For example, from the verbal stem ነገር /*nəgəri*/, it is possible to generate the following words: ነገርኳቸው /*nəgərikwatʃəwi* 'I told them'/, ተነገረች /*tənəgərətʃi* 'she has been told'/, ነገረኝ /*nəgərəɲi* 'he told me'/, ነገርሽው /*nəgəriʃiwi* 'you told him'/, etc. Words generated from stems could have both prefixes and suffixes. For example, the word ስላልተደመለጓቸውና /*silalitədəwələlatʃəwina*/ has three prefixes (ስለ /*silə* preposition/, አል /*ʔəli* negation/, ተ /*tə* passivizer/), a perfective stem ደመል /*dəwəli*/ and four suffixes (ኧ /*ə*/, ል /*li* beneficative/, አቸው /*ʔətʃəwi* [3rd person, plural]/, ና /*na* conjunction/). Many words are also formed through inflectional process. Amharic nouns and adjectives are inflected with number, person, gender, definiteness, etc. to form many words in surface forms. Some Amharic functional words such as prepositions, conjunctions, etc. also undergo morphological process to form variants. They exist either by merging with each other or by affixation with other words. For example, from the word ሌላ /*lela* 'other'/ variants that could be generated include ሌላዎች /*lelawətʃi*/, የሌላ /*jəlela*/, ከሌላ /*kəlela*/, በሌላ /*bəlela*/, የሌሎች /*jəlelotʃi*/, ስለሌሎቻችን /*silələlotʃatʃini*/, ሌላው /*lelawi*/, etc.

In general, the formation of thousands of words, especially from verbal roots makes the analysis, annotation and tagging of Amharic text a difficult task. This level of morphological complexity has significantly contributed to the difficulty of producing linguistic resources and NLP applications for Amharic. Therefore, the morphological structure of Amharic is an important issue while developing Amharic NLP tools, resources and applications.

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 relatively a long history [37–39]. Musaid [13] investigated the effectiveness of word-based, stem-based, and root-based representation of documents and queries. The word-based and stem-based text representations miss relevant documents while root-based representation retrieves non-relevant documents. Moukdad [8] compared the effects of stem and root on Arabic IR. The results of their experiments indicate that the use of stem is more effective than root. Larkey et al. [37] investigated the effects of light stemming (removal of prefix and suffix) on Arabic IR. A comparison was made between stem-based and root-based retrieval. The finding indicates that light stemmer outperforms root analyzer and other stemmers which are based on detailed morphological analysis. Abdusalam [40] presented an Arabic text retrieval technique using lexicon-based light stemming. The study evaluated the effectiveness of lexicon-based light stemming, Arabic patterns, root, expanding query and filtering foreign words using n-grams on TREC 2001 test collection. According to the results, the preprocessing techniques such as normalization, stopword removal and light-stemming improve retrieval results whereas n-grams and roots decrease the performance. Al-Hadid et al. [41] developed a neural network-based model where documents and queries are represented using stems and their similarity is computed using cosine similarity. The lexicon-based stemming and the relevance feedback approaches perform better than light-stemming approach alone. Ali et al. [42] 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 another Semitic language, spoken mainly in Israel. Carmel [43] presented a morphological disambiguation tool 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 and the performance of the search engine increased. Ornan [44] designed Hebrew search engine by applying a rule-based morphological analysis. The design of the search engine considers the construction of morphological, syntactic and semantic analyzers. The search engine eliminates words unsuited both to the syntax and the semantics 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 [25] studied the retrieval effectiveness of word-based, stem-based, and root-based text representations on Amharic language. The experiments were carried out by running 40 queries on 548 documents using OKAPI system and the study concludes that stem-based retrieval is slightly better than root-based. Mindaye et al. [26] developed an Amharic search engine using stems. The system was tested with 11 queries on 75 news documents. On this small

collection, the average precision and recall values they report are 0.65 and 0.95, respectively, using OR operator in between query terms, and 0.99 and 0.52, respectively, for AND operator. Argaw et al. [45] developed dictionary-based Amharic-English IR system. Documents and queries are represented using bag-of-words. In their work, stopwords are removed using Inverse Document Frequency (IDF) and stopwords list. The average precisions of 0.3615 and 0.4009 were achieved using IDF and stopword list, respectively. Argaw et al. [46] also have developed dictionary-based Amharic-French IR system with and without word sense discrimination using bag-of-words approach. Word sense discrimination refers to the distinction between different word meanings without sense assignment [47]. Stemming was applied to remove prefix and suffix. The experiments were conducted on the Swedish Institute of Computer Science (SICS) and Lucene search engines. Stopwords were removed by using IDF. The authors presented that the results are better when using SICS than Lucene. The word sense discrimination performs slightly better than non-discrimination.

The effect of query expansion using word embedding was evaluated in some researches. The impact of query expansion on IR based on NLP tools and Word2Vec embedding algorithms was verified in [48]. Continuous Bag of Word (CBOW) and Skip-gram models were trained on TREC corpus and used to select semantically related terms to the initial query terms. In order to obtain better retrieval result, the two word embedding models were combined with stemming. Okapi BM25 probabilistic weighting scheme with default parameters was used for ranking retrieved documents. The experiment was conducted by running 72 queries. The mean average precision (MAP) 0.1769 and 0.128 are obtained for Skip-gram and CBOW models, respectively. Whereas the Normalized Discount cumulative Gain (NDCG) values 0.4576 and 0.3841 are obtained for skip-gram and CBOW models, respectively. The use of local and global word embedding in query expansion for *adhoc* IR was studied in [49]. The idea was to evaluate the effect of the local word embedding CBOW and the global word embedding GloVe on query expansion. Various experiments were conducted on TREC12, Wiki, news, web, and robust corpora after stemming using Language Modeling (LM) approach. The result indicates that query expansion using local embedding is better than global embedding for IR. However, the retrieval effectiveness of query expansion using local embedding is poor. Amharic semantic-based IR system was developed using BM25 model [50]. Documents and queries were processed by a stemmer. The retrieval effectiveness of the system was evaluated by running 10 queries on 8759 documents and performed an average recall and precision of 0.84 and 0.23, respectively. Hadi et al. [51] developed Arabic document retrieval system based on word-embedding and prospect-guided as query expansion techniques. The deep averaging network in vector space model, the probabilistic Okapi BM25 model and the two representative word embedding methods for automatic query expansion was used. Various experiments were carried on TREC test collections. Stemming was performed and has shown a significant impact on the performance of Arabic text retrieval. Their finding indicates that deep average network could not improve performance using only BM25. However, the retrieval performance outperformed all baselines after integrating the two query expansion techniques.

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 such as English. This has not produced the desired retrieval result as documents could not be represented appropriately. Only few studies have tried to consider the issue of document representation in a systematic way [13]. In our approach, we address this crucial issue of document representation in the development of Amharic IR.

4. Amharic IR System Design and Text Representation

The main objective of this work is to identify the optimal representations for documents and queries in Amharic IR. Our approach focuses on the selection of the term structures and stopwords, based on the morphological characteristics of the language. Taking these issues into consideration, we also proposed an Amharic retrieval system as shown in Figure 1, which is slightly different from the basic architecture of IR systems. In the architecture

we propose, stopwords are removed after the application of morphological analysis on documents and queries. Both documents and queries pass through the same preprocessing and linguistic text processing modules.

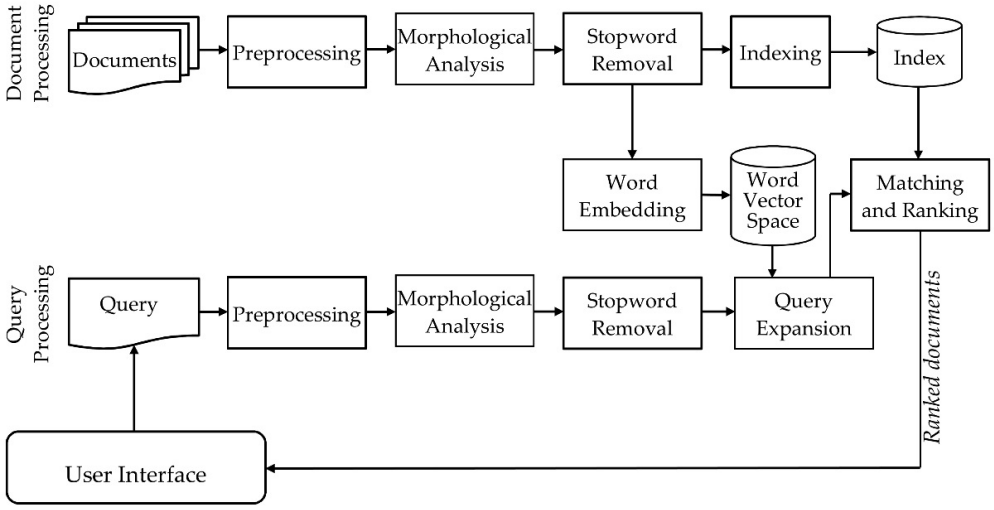


Figure 1. Architecture of the proposed Amharic IR system.

4.1. Text Preprocessing

The preprocessing involves removal of tags and punctuation, language specific tokenization, and character normalization. Punctuation marks are removed as they do not bear content relevant to IR system. Character normalization is required due to the fact that some characters have similar pronunciation but different graphemes. The characters ሐ /hə/, ኀ /hə/, ኸ /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 ዐ /ʔə/ and its modifications are normalized to their corresponding modifications of አ /ʔə/. The fourth orders ሃ /ha/, ሓ /ha/, ኃ /ha/ and ኻ /ha/ are normalized to ሀ /hə/ whereas the fourth orders ኣ /ʔa/ and ዓ /ʔa/ are normalized to አ /ʔə/. Thus, ስለጸሀፊዋ /sɨlətsʼəhafiwa 'about the secretary'/ can also be written as ሥለጸሀፊዋ /śɨlətsʼəhafiwa/, ስለጸሓፊዋ /sɨlətsʼəhafiwa/, ሥለጸሓፊዋ /śɨlətsʼəhafiwa/, ሥለጸኀፊዋ /śɨlətsʼəhafiwa/, ስለጸኀፊዋ /sɨlətsʼəhafiwa/, ሥለጸሐፊዋ /śɨlətsʼəhafiwa/, etc. although some of them rarely appear in a text.

4.2. Morphological Analysis

Morphological analysis is required for stemming and lemmatization where both aim to conflate variants of words into a unique unit, a stem or a root. These text units are then used to represent documents and queries for document indexing and query/document matching. Stemming is usually applied since lemmatization is more computationally consuming for just slight effectiveness improvements in IR tasks [52]. Stemming has also been applied in Amharic IR systems [25,26]. In our work, we make the hypothesis that stemming is insufficient for Amharic and that more sophisticated text analysis should be used because of the complexity of the language.

Morphological variants of Amharic, especially verbs can have more than one stem types. From a given Amharic root, more than 10 basic stems could be generated [53]. For example, the morphological variants ነገረ /nəgərə 'he told'/, ተናጋሪ /tənagari 'orator'/, and አናገረ /ʔənagərə 'he communicated'/ have the basic stems ነገር- /nəgəri-/, ናጋር- /nagari-/ and ናገር /nagəri-/, respectively. As a result, stemming provides different stems though the word variants are semantically similar, which means that Amharic verbal stems need one more reduction analysis for extracting words' roots. Indeed, verbal stems are formed from roots and all variants of an Amharic verb have one common root. For example, the common

root for the aforementioned examples of morphological variants is ንግር /*n-g-r*/. Therefore, roots are more appropriate than stems for Amharic as roots conflate word variants more accurately. Thus, we developed a new root-based representation for text representation for Amharic IR. We test our hypothesis experimentally by considering stem-based and root-based morphological analyses where extracted basic stems and roots of words from documents and queries. Basic stems serve for the formation of derived stems and surface forms of words. In Amharic, basic stems are usually derived from roots by inserting vowels between radicals.

4.2.1. Stem-Based Morphological Analysis

We created the stem-based index using the basic stems of words. Several words can be formed by attaching affixes to stems. Therefore, we performed morphological analysis for extracting stems from the rest of morphemes. We conflated variants of words to their basic stems. For example, the morphological analysis used to extract the stems of the words አልተመለሱም /*ʔəlitəmələsumi* 'they have not returned'/ and ከቤተሰቦቹ /*kəbetəsəbotʃu* 'from his families'/ are shown as follows (see Notations at the end of this article).

አልተመለሱም	ከቤተሰቦቹ
አል_ተ_መለስ_ኡ_ም	ከ_ቤተሰብ_ኦች_ኡ
ʔəli_tə_mələsi_ʔu_mi	*kə_betəsəbi_ʔotʃi_ʔu*
[*neg*]-[*pas*]-[*stem*]-[*3,p*]-[*ncmp*]	[*pre*]-[*stem*]-[*p*]-[*3,s,m*]

Furthermore, basic stems are used to form derived stems, which in turn are used for the formation of surface forms of words. The derived stems include causative (አ- /*ʔə-*/ and አስ- /*ʔəsi-*/), passive (ተ- /*tə-*/), infinitive (መ- /*mə-*/), and reduplicative types of verbal stems. For example, the variants ተደወለ /*tədəwələ* 'is called'/, ከመደወሏ /*kəmədəwəlwa* 'as soon as she called'/, በአስደወለ /*bəʔəsidəwələ* 'since he caused to call'/, and ሲደዋወሉ /*sidəwawilu* 'as they called each other'/ have the derived stems ተደውል- /*tədəwili-*/, መደወል- /*mədəwəli-*/, አስደወል- /*ʔəsidəwəli-*/ and ደዋወል- /*dəwawəli-*/, respectively and a basic stem ደወል- /*dəwəli-*/. With regard to their meaning, there is no conceptual difference between derived and basic stems. Moreover, the origins of derived stems are basic stems, and basic stems are the shortest and the most common stems for many variants (see Table 3 where core meaning is 'kill'). For the stem-based indexing and retrieval, we represent the variants of words by their basic stems.

Table 3. Derived and basic stems in Amharic.

Variant	Derived Stem	Basic Stem
ከተገዳይ /*kətəgədaji*/	ተገዳል- /*təgədali-*/	ገዳል- /*gədali-*/
ስላጋደለ /*silagadələ*/	ኢጋደል- /*ʔəgadəli-*/	ጋደል- /*gadəli-*/
ካስገደለ /*kasigədələ*/	አስገደል- /*ʔəsigədəli-*/	ገደል- /*gədəli-*/
ከገዳደለ /*kəgədadələ*/	ገዳደል- /*gədadəl-*/	ገደል- /*gədəli-*/

Although basic stems are better to conflate more variants than derived stems, more than one type of basic stems exist for variants (see Table 3). Therefore, in case of verbs, it is impossible to conflate Amharic variants even using basic stems. However, variants of primary nouns, adjectives, and adverbs have one common basic stem. Therefore, verbs and words derived from verbs need further morphological analysis to be represented by a common form. The morphological analysis of some words requires palatalization to extract basic stems. This has been achieved after separating -ኡ and -ኢያ from a stem. For example, the morphological analyses of the words የጎጂዎቹ /*jəgodʒiwotʃi* 'of harmful [*p*]'/, ገዳዮች /*gədajotʃi* 'killers'/ and መጨረሻ /*mətʃʼərəʃa* 'end'/ are presented as follows.

የጎጂዎቹ	ገዳዮች	መጨረሻ
የ_ጎድ_ኡ_ዎች	ገድል_ኡ_ኦች	መ_ጨረስ_ኢያ
jə_godi _ ʔi_wotʃi	*gədili _ ʔi_ʔotʃi*	*mə_tʃʼərəsi_ʔija*
[*gen*]-[***stem***]-[*pal*]-[*p*]	[***stem***]-[*pal*]-[*p*]	[*inf*]-[***stem***]-[*pal*]

4.2.2. Root-Based Morphological Analysis

Roots are the bases for the formation of verbal stems and many Amharic words as the origin of verbs and words derived from the verbal stems is root. Stem and root have the same form for primary nouns, adjectives, adverbs, and functional words. For Amharic nouns, adjectives, and adverbs derived directly from verbal roots and stems, we proposed the roots of their corresponding verbs as index and query terms. For example, the morphological analysis of the verb ከሰበርኳቸውማ */kəsəbərikwatʃəwima* 'if I break them even'/ and the derived noun ስብራቴ */sibirate* 'my state of being broken'/ are presented below.

ከሰበርኳቸውማ	ስብራቴ
ከ_ሰበር_ኩ_አቸው_እማ	ስበር_አት_ኤ
kə_səbərɨ_ku_ʔətʃəwɨ_ʔɨma	*sɨbɨrɨ_ʔətɨ_ʔe*
[*pre*]-[***stem***]-[*1,s*]-[*3,p*]-[*foc*]	[***stem***]-[*nom*]-[*1,s*]
ከ_ስ-ብ-ር_ኩ_አቸው_እማ	ስ-ብ-ር_አት_ኤ
[*pre*]-[***root***]-[*1,s*]-[*3,p*]-[*foc*]	[***root***]-[*nom*]-[*1,s,pos*]

Adjectives derived from primary nouns are represented using the root of the corresponding nouns. Nouns derived from primary adjectives are also represented using the root of the corresponding adjectives. The root representation of primary nouns, adjectives, adverbs and functional words is different from verbal root representation. For example, the root of the noun መኪናዋ */məkinawa* 'her car'/, ደጋግ */dəgagɨ* 'generous'/, and ሌሎች */lelotʃɨ* 'others'/ are መኪና */məkina*/, ደግ */dəgɨ*/, and ሌላ */lela*/, respectively. However, the root of the verbals such as መልስ */məlisɨ* 'answer'/ and ረዥም */rəʒɨmɨ* 'long'/ are ም-ል-ስ */m-s-l*/ and ር-ዝ-ም */r-z-m*/, respectively. The reduced form of some variants of a verb is represented by the corresponding radical form. For instance, the root of the verb ሞተ */motə* 'he died'/ and ኖረ */norə* 'he lived'/ are ም-ው-ት and ን-ው-ር, respectively. Morphological variants of Amharic words, especially verbs can have more than one stem, but still a common root. All variants of an Amharic verb can thus be represented by a single root during indexing. On the other hand, semantically unrelated words hardly ever have a common root. To sum up, basic stem text representation is robust to represent primary noun, adjectives and adverbs, and functional words whereas root is robust to represent all types of words, including verbs.

4.3. Amharic Stopword Identification and Removal

One of the major preprocessing tasks in IR and many other text processing applications is stopword removal. Accordingly, stopword lists have been constructed for many languages. However, standard stopword list is unavailable for Amharic IR yet. The common trend for identifying and removing stopwords is to do it before applying morphological analysis on words in a text. This is also what has been carried out in the previous Amharic IR studies. We think this is an inappropriate way to consider stopwords for Amharic. Some Amharic stopwords do not necessarily exist as standalone words and others may appear with other words as prefix and suffix. For example, 'the' is usually considered as a stopword in English; its Amharic equivalent is a suffix '-ኡ */-ʔu*/' or '-ው */-wi*/' that does not appear as a standalone word. Accordingly, 'the house' and 'the student', for instance, are equivalent to 'ቤት */beti*/+ -ኡ */-ʔu*/' → 'ቤቱ */betu*/' and 'ተማሪ */təmari*/+ -ው */-wi*/' → 'ተማሪው */təmariwi*/', respectively. Terms can appear in various morphological structures as there can be several sequences of affixes representing articles, prepositions, numbers, etc. For instance, the stopword ውስጥ */wisɨtʼɨ* 'in'/ has the following variants: ውስጣዊ */wisɨtʼawi*/, ውስጣችን */wisɨtʼatʃɨni*/, ውስጥና */wisɨtʼɨna*/, ውስጥም */wisɨtʼɨmi*/, የውስጥ */jəwisɨtʼɨ*/, ለውስጥ */ləwisɨtʼɨ*/, በውስጥ */bəwisɨtʼɨ*/, ከውስጥ */kəwisɨtʼɨ*/, የውስጥና */jəwisɨtʼɨna*/, etc. Furthermore, some stopwords merge with each other or other words to form new words. Thus, it is impossible to find and remove most Amharic stopwords before the application of morphological analysis. This calls for a different consideration for Amharic stopword identification and removal in comparison with morphologically simpler languages such as English.

As we designed stem-based and root-based Amharic IR system, we also constructed stopword lists based on stem and root forms. In either of the cases, stopwords are identified

after morphological segmentation of words from a large corpus of documents representing various domains and sources. For example, words such as ስላላመጣቸው· */silalaməṭ'aṭʃəw* 'since he did not bring them'/ and ከትልልቆችም */kətililik'otʃim* 'even from big ones'/ undergo the following morphological process to extract attached stem-based stopwords.

ከትልልቆችም	ስላላመጣቸው·
ከ_ትልቅ_አች_ም	ስለ_አል_አ_መጥ_አቸው·
kə_tilik'_otʃ_im	*silə_ʔəli_ʔə_məṭ'_ʔəṭʃəw*
[*pre*]-[***stem***]-[*p*]-[*foc*]	[*com*]-[*neg*]-[*cau*]-[***stem***]-[*3,p*]

Similar to the stem-based stopword list, the root-based stopwords were built based on the root-based morphological process as shown in the following example.

ከትልልቆችም	ስላላመጣቸው·
ከ_ት-ል-ቅ_አች_ም	ስለ_አል_አ_ም-ጥ_አቸው·
kə_t-l-k'_otʃ_im	*silə_ʔəli_ʔə_m-t'_ʔəṭʃəw*
[*pre*[-[***root***]-[*p*]-[*foc*]	[*pre*]-[*neg*]-[*cau*]-[***root***]-[*p*]

Statistical information about terms plays a significant role for identification of stopwords. However, the notion of term depends on the characteristics of languages. For morphologically simple languages such as English, stems can be considered as terms. However, this is not exactly the case with morphologically complex languages such as Amharic. We hypothesize that morphemes used to form Amharic words can be used as a basis for computing term statistics. Thus, in this work, we consider morphemes as terms. Both the stem-based and root-based stopwords were created based on the aggregation of the morpheme statistical information (frequency, mean, variance and entropy) from Amharic corpus. Accordingly, for the entire documents that we collected, we compute frequency, mean, variance and entropy of each morpheme in order to generate corpus-based Amharic stopword list as detailed below.

The frequency of morpheme is represented by document frequency and collection frequency. The document frequency of a morpheme is the number of documents where the morpheme occurs whereas the collection frequency is the total morpheme frequency throughout the corpus. In this work, we used un-normalized morpheme frequency. Morphemes were ranked according to their document frequency and collection frequency. Then, a threshold value was used to determine stopwords. Morphemes that are evenly distributed throughout the collection and satisfy the threshold value are considered as stopwords. The document frequency df of each morpheme is computed as:

$$df(M_i) = \sum_{i=1}^{N} morpheme\_status(D_i) \tag{1}$$

where M_i is the ith morpheme in the corpus, D_i is the ith document in the corpus, N is the total number of morphemes in the collection. If a morpheme appears in a given document, its status is 1 otherwise 0. The collection frequency cf of each morpheme is computed as:

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

where MFD_i is the morpheme frequency in each document, N is the total number of documents in the corpus.

The mean value of each morpheme is used to measure the overall distribution of morphemes in the whole corpus. The mean probability ***mp*** of each unique morpheme in all documents is computed as:

$$\boldsymbol{mp}(M_i) = \frac{\sum_{i=1}^{N} p(M_i)}{N} \tag{3}$$

where N is total number of documents and $\boldsymbol{p}(M_i)$ is morpheme probability which is computed as:

$$\boldsymbol{p}(M_i) = \frac{MF}{TM} \tag{4}$$

where MF is morpheme frequency in each document and TM is the total number of morphemes in a document.

The variance of morphemes is used to check the distribution of morphemes throughout the documents in the corpus. Variance v is computed as:

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

where $v(M_i)$ is the ith morpheme variance, N is the total number of distinct morphemes in the document, $n(M_i)$ is normalized morpheme frequency in a document and $m(M_i)$ is mean value computed as follows.

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

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

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

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

where $\boldsymbol{p}(M_i)$ is the probability of morpheme frequency and is calculated by dividing the morpheme frequency with the total number of morphemes in the document.

The stopwords were selected based on the aggregation of document frequency, mean, variance and entropy values of morphemes. Initially, four lists each containing the top 250 morphemes were selected using the statistical information of morphemes in the corpus. Out of these lists, 180 morphemes located across all the four lists were then selected through empirical analysis (see Table 4).

The final stopword list also includes few words which were selected manually based on Amharic subword class criteria. These words are functional words used mainly for the formation of phrases, sentences and paragraphs. These words are characterized by the lack of meaning by their own, inability to undergo morphological derivation and inflection, lack of morphemes for various parameters, and they have small word size [53]. It includes words such as ወደ /*wədə* 'towards'/, እንደ /*ʔɨndə* 'like'/, ስለ /*sɨlə* 'about'/, እስከ /*ʔɨskə* 'up to'/, ወዘተ /*wəzətə* 'and so on'/, ጎሽ /*goʃ* 'bravo'/, ዋ /*wa* 'warning'/, ይልቅ /*jɨlik* 'instead'/, etc. Based on this criterion, we selected 42 words. Thus, the final stopword list contains 222 terms.

The stem-based and root-based stopwords are then removed from the respective documents and queries during the process of indexing and query processing. The stem-based and root-based stopword lists we have built are re-usable, and they are one of the outputs of this research as we used a large set of documents from different domains and sources.

4.4. Indexing

For the purpose of IR, index terms are considered based on their frequency. Computing the frequency of Amharic terms is not a straightforward as there are several words derived from the same form. Root–based representation conflates all variants to a common root whereas stem-based representation does not guarantee this. The root-based representation can thus better compute the actual term frequency. The stem-based representation computes term frequency inaccurately for words derived from the same form. For example, variants

of ስ-ብ-ር /*s-b-r*/ are represented by more than one term for word-based and stem-based representations whereas a single term represents them in root-based representation (see Table 5 in which the frequency are based on the document presented in Appendix A).

Table 4. Sample bound and free morphemes with highest statistical values.

Morpheme	Meaning	Rank			
		Frequency	Mean	Variance	Entropy
የ /*jə*/	of	1	1	1	1
በ /*bə*/	by	2	2	2	2
ኡ /*ʔu*/	they	3	3	3	3
ኧ /*ʔə*/	he	4	4	5	17
ው /*wi*/	the	5	5	6	4
አል /*ʔəli*/	not	6	9	15	9
ኦች /*ʔotʃi*/	many	12	12	10	12
እና /*ʔina*/	and	13	14	14	13
ላይ /*laji*/	on	29	31	37	29
ድ-ር-ግ /*d-r-g*/	act	31	33	38	31
ነው /*nəwi*/	is	41	43	52	40
ህ-ው-ን /*h-w-n*/	happen	55	21	23	20
ወደ /*wədə*/	towards	71	74	80	67
እንደ /*ʔinidə*/	like	22	27	36	19
ያ /*ja*/	that	25	26	27	25

Table 5. Frequency of terms derived from ስ-ብ-ር /*s-b-r*/ in a sample document.

Variant	Frequency		
	Word-Based	Stem-Based	Root-Based
የስብራት /*jəsibirati*/	1	15 (ስብር /*sibiri*/)	26 (ስ-ብ-ር /*s-b-r*/)
ስብራቶች /*sibiratotʃi*/	2		
ስብራትን /*sibiratini*/	2		
ስብራትና /*sibiratina*/	1		
ስብራት /*sibirati*/	7		
ስብራቱ /*sibiratu*/	2		
የተሰበረውን /*jətəsəbərəwuni*/	1	7 (ሰበር /*səbəri*/)	
መሰበርን /*məsəbərini*/	1		
መሰበርና /*məsəbərina*/	1		
መሰበር /*məsəbəri*/	4		
መሰባበር /*məsəbabəri*/	1	1 (ሰባበር /*səbabəri*/)	
ሰባራ /*səbara*/	3	3 (ሰባር /*səbari*/)	

Generally, the term frequency for root-based representation is higher than or equal to the stem-based one; so will be the associated with if computed with the usual formula (See Equation (8)).

$$w(t,d) = tf.log\frac{N}{df} \quad (8)$$

where $w(t,d)$ is weight of a term t in a document d, tf is term frequency, N is total number of documents in a corpus, and df is document frequency. On the other hand, the number of stem-based terms is larger than or equal to the number of root-based terms in Amharic documents. For example, in the document presented in Appendix A, there are 405 words, 213 stems and 168 roots. The effect of morphological analysis on Amharic IR using stem-based and root-based indexes is evaluated in Section 5.

4.5. Word Embedding

Similar to other languages, in Amharic language semantically related words co-occur many times in a text document. User information need might be incomplete or contain all co-occurrence words. As a result of this, relevant documents could not be retrieved. Therefore, user information need should be completed by query expansion in order to retrieve more relevant documents. To overcome the problem of vocabulary mismatch between index term and query term, we used word embedding vector representation for each word in the vocabulary. Each query terms are expanded by semantically and syntactically related terms using word2vec developed by Mikolov et al. [54]. To verify the impact of query expansion on Amharic IR using word embedding algorithms, we created CBOW and Skip-gram models on stem-based and root-based corpora. CBOW and skip-gram models capture order relationships between words and they are effective for NLP tasks involving the use of word similarity and word analogy. We used them to capture semantic relations between words in queries and relevant documents. Four neural network models: stem-based with CBOW, stem-based with Skip-gram, root-based with CBOW and root-based with Skip-gram models are created offline and used for expanding stem-based and root-based query terms using stem-based and root-based corpora, respectively. The original query terms are matched with vocabulary words in the models to find their vector values. Then, the similarity between the features of each original query word and vocabulary words is computed to identify similar terms for each query term. The similarity *sim* between a query term q and a corpus word d is computed using cosine similarity as shown in Equation (9).

$$sim(q,d) = \frac{\sum_i qi.di}{\sqrt{\sum_i qi^2 \sum_i di^2}} \quad (9)$$

where qi is vector representation of the ith query term and di is the vector representation of the ith word in the corpus. The top 5 most related terms are used to expand query terms. Nearest neighbors or syntactically related words are used to expand query. Words that are used in similar contexts with respect to a given word frequently are used as expanded terms.

4.6. Matching and Ranking

In the proposed system, the extraction of index and query terms from documents and user queries uses the same workflow with text preprocessing, morphological analysis and stopword removal. Searching for relevant documents is based on matching query terms (representing information need of users) with index terms (representing documents). We used exact vocabulary term matching which searches documents that contain the query terms without analyzing the semantics of words and without considering the semantic connections between them. The retrieval probability of a relevant document for a given query is different in case of stem-based and root-based retrievals. In general, the retrieval probability of a relevant document for a given query in case of root-based matching is higher than stem-based matching. For example, consider the sample document (Appendix A) and the following four query terms derived from ስ-ብ-ር /*s-b-r*/:

1. ስብራት /*sibirati* 'the state of something which is broken'/
2. መሰበር /*məsəbəri* 'to break'/
3. ሰባራ /*səbara* 'something which is broken'/
4. አሳባሪ /*ʔəsabari* 'cause to break'/.

The document (Appendix A) is relevant to the four queries. The retrieval probability of the document in the case of stem-based matching is 0.032 for term query (1), 0.017 for (2), 0.004 for (3) and 0 for (4). However, in case of root-based matching, the retrieval probability of the document (at Appendix A) to the four queries is similar (0.057), which is better than stem-based matching. Furthermore, the stem-based text representation retrieves non-relevant documents since it may conflate semantically unrelated words.

We use Lemur toolkit for ranking. For a given query Q and a collection of retrieved documents D, the Lemur toolkit ranks retrieval results based on their possible relevance. It implements both BM25 and language modeling, where the document length is considered. BM25 ranks documents based on the following equation:

$$Score\ (D,Q) = \sum_{i=1}^{n} IDF(qi).\frac{f(qi,D).(k1+1)}{f(qi,D)+k1.(1-b+b.\ \frac{|D|}{avgdl})} \quad (10)$$

where f (qi,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 whereas k_1 and b are free parameters. IDF (q_i) is the inverse document frequency weight of the query term q_i. From Equation (10), $Score$ (D,Q) would be higher when query terms in Q have higher frequencies in document D. For each qi in Q, the following inequality holds true for Amharic documents.

$$f_r(qi,D) >= f_s(qi,D) >= f_w(qi,D) \quad (11)$$

where f_r, f_s, and f_w denote term frequency in root-based, stem-based and word-based representations, respectively. Thus, it can be inferred that the root-based representation of queries and documents provide better information for ranking as $f_r(qi,D)$ provides the highest possible score.

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 Kullback-Leibler (KL) divergence ranking function captures the term occurrence distributions and computed as:

$$\mathrm{KL}\ (Q\theta, D\theta) = \sum_{w \in V} p(w|Q\theta) \log \frac{p(w|Q\theta)}{p(w|D\theta)} \quad (12)$$

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.

5. Experiment

Since Amharic IR test collection was unavailable, the first task of this research was to create such resource. We followed the TREC approach so that the general TREC evaluation framework could also be followed. We also compared the performance of the proposed system with that of previous studies and Google Amharic.

5.1. Corpora Preparation

5.1.1. Test Collection

Test collection plays a crucial role for designing, developing, and maintaining IR systems. However, test collection for Amharic *adhoc* IR did not exist. Thus, we created the first Amharic IR test collection named 2AIRTC and share it to the IR research community [50]. The test collection was created based on the framework used in TREC and Cranfield project.

The Amharic test collection contains representative documents and topics. The test collection consists of 12,583 documents, 240 topics and the corresponding relevance judgments. Documents were collected from Web, news agencies, and individuals. The topic set was created by considering both current issues and the topics that were likely to be treated in our initial corpus. They cover diverse issues and include short and medium topics in addition to collocation. The document relevance judgment was made manually by assessors using a precise guideline. We ran the title fields of the topics on our initial corpus using Lemur toolkit and Web using Google search engine to acquire a first retrieved document list of a maximum of 50 documents per topic. We fused both retrieved document list (Lemur LM and Google retrieval results) on our initial collection and Web documents for the complementary documents). Each document from the fused list was then judged

for relevance. Furthermore, exhaustive relevance judgments were performed on some topics on the document collection to acquire more relevant documents. A document is marked as relevant based on the narrative information in the topic; thus, it should not simply contain words from a query but rather fulfill the information need. The summary of the test collection is presented in Table 6 and more details are available in [50].

Table 6. 2AIRTC relevance judgment statistics.

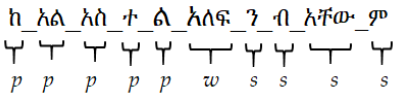
Parameter	Size
Total number of topics	240
Average number of relevant documents per topic	22
Minimum number of relevant documents per topic	10
Maximum number of relevant documents per topic	172

5.1.2. Morphologically Annotated Corpora

Annotated corpora play significant role in the development of NLP tools and the evaluation of tasks such as IR. Morphological analysis is one of the fundamental NLP process to derive stem, root and grammatical parts of words based on its internal structure [21]. This could be carried out manually, automatically by NLP tools such as a morphological analyzer, or semi-automatically. The existing Amharic morphological analyzers are limited in scope and function, and are unsuitable for query and document processing. We created stem-based and root-based morphologically annotated corpora using semi-automatic method. These documents are part of our 2AIRTC collection. Each word in each document in the corpus are morphologically segmented into affixes and base (basic stems or roots). The stem and root extraction were performed by removing character(s), changing the shape of character, or simply segmenting stem from the rest of characters (affixes). Different word classes have different word structures during word formation, but many of them have prefix(s)-stem, stem-suffix(s), or prefix(s)-stem-suffix(s) word formation structures. The number of prefixes and suffixes varies from one to five [16]. The general structure of a morphologically annotated word *W* is:

$$[p\_]^* \; w \; [\_s]^*$$

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 verb ካላስተላለፍንባቸውም */kalasɨtəlaləfinɨbatʃəwɨmɨ* 'if we did not transfer for them [*foc*]'/ is morphologically segmented as follows.



The first five morphemes are prefixes; አለፍ */ʔələfɨ/* is the basic stem whereas the last four morphemes are suffixes.

Stem-based corpus annotation: Majorities of Amharic words are composed of stems and attached affixes. Therefore, the stem-based morphological annotation segments word forms into more general representation known as basic stem and affixes. For example, the verb አልተለማመደም */ʔəlɨtələmaməd*ə*mɨ* 'he did not practice'/, ረዣዥሞቹ */rəʒaʒɨmotʃu* 'the longest ones'/, the adjective የበለሆችን */jəbɨlɨhotʃɨnɨ* 'of intelligent [*acc*]'/, the verb በመሆኑ */bəməhonu* 'since it happened'/, the noun በገንዘቦቻንና */bəgənɨzəbatʃɨnɨna* 'through our money'/, and the functional word የሌሎቻቸውን */jəlelotʃatʃɨhunɨ* 'of other [*acc*]'/ are annotated as indicated below.

አልተለማመደም	የበልሆችን	በመሆኑ
አል_ተ_ለመድ_ኧ_ም	የ_በልህ_አች_ን	በ_መ_ሆን_ኡ
ʔəli_tə_ləmədi_ə_mi	*jə_bilihi_ʔotʃ_i_ni*	*bə_mə_honi_ʔu*
*[neg]-[pas]-[**stem**]-[3,s,m]-[ncom]*	*[gen]-[**stem**]-[p]-[acc]*	*[pre]-[inf]-[stem]-[def]*
በገንዘባችንና	የሌሎቻችሁን	ረኀቦሞቹ
በ_ገንዘብ_አች_ና	የ_ሌላ_አች_አችሁ_ን	ረኀቦም_አች_ኡ
bə_gənizəbi_ʔətʃini_na	*jə_lela_ʔotʃi_ʔətʃihu_ni*	*rəʒimi_ʔotʃi_ʔu*
*[pre]-[**stem**]-[1,p,pos]-[conj]*	*[gen]-[**stem**]-[p]-[2,p,pos]-[acc]*	*[**stem**]-[p]-[def]*

Root-based corpus annotation: Amharic roots are usually the base of stems and surface forms of words whereas stem is the base of surface forms of words. Thus, the origins of Amharic words are root rather than stems. Therefore, to analyses the impact of root for Amharic IR we created the root-based morphological annotated corpus which is part of 2AIRTC. Each word in each document is annotated morphologically for segmenting the root of each word from the rest of morphemes in a word. Therefore, the root-based corpus contains roots of words and their affixes. For instance, the verb ስላገዳደለቻቸው */sɨlagədadələcatʃəw* 'since she makes them to kill each other'/ is annotated as 'ስለ_አ_ግ-ድ-ል_ኧች_አቸው */sɨlə_ʔə_gɨdɨl_ətʃi_ʔətʃəw/*'. ስለ */sɨlə* 'since'/ is preposition, አ */ʔə/* is causative, ግ-ድ-ል */g-d-l* 'kill'/ is a root, ኧች */ətʃi/* and አቸው */ʔətʃəw/* are person markers. Similarly, other word classes are annotated. For example, the corresponding root-based annotations of sample words for which stem-based annotations are shown above are presented as follows.

አልተለማመደም	የበልሆችን	በመሆኑ
አል_ተ_ል-ም-ድ_ኧ_ም	የ_በልህ_አች_ን	በ_መ_ህ-ው-ን_ኡ
ʔəli_tə_l-m-d_ə_mi	*jə_bilihi_ʔotʃi_ni*	*bə_mə_h-w-n_ʔu*
*[neg]-[pas]-[**root**]-[3,s,m]-[ncom]*	*[gen]-[**root**]-[p]-[acc]*	*[pre]-[inf]-[**root**]-[def]*
በገንዘባችንና	የሌሎቻችሁን	ረኀቦሞቹ
በ_ገንዘብ_አች_ና	የ_ሌላ_አች_አችሁ_ን	ር-ዝ-ም_አች_ኡ
bə_gənizəbi_ʔətʃini_na	*jə_lela_ʔotʃi_ʔətʃihu_ni*	*r-ʒ-m_ʔotʃi_ʔu*
*[pre]-[**root**]-[1,p,pos]-[conj]*	*[gen]-[**root**]-[p]-[2,p,pos]-[acc]*	*[**root**]-[p]-[def]*

The root representation of words derived from verbal root is slightly different from non-verbal root. Roots of verbal are represented by radicals but might not in case of non-verbal roots. On the other hand, the stem and root of non-verbs have similar representation. However, the stems and roots of verbs are different. The details of them are presented in Table 7.

Table 7. Examples of stem and root representation.

Type of Morphology	Sample Word	Meaning	Stem	Root
Derived from verbal root	ስንፈት */sɨnifətɨ/*	weakness	ስንፍ	ስ-ን-ፍ
	ጠንካራ */t'ənikara/*	strong	ጠንካር	ጥ-ን-ክ-ር
	ነገረኝ */nəgərəɲɨ/*	he told me	ነገር	ን-ግ-ር
	ቀላጅ */k'əladʒɨ/*	comedian	ቀላድ	ቅ-ል-ድ
	ሟች */mwatʃɨ/*	the one who died	ሟች	ም-ው-ት
Not derived from verbal root	እግረኛ */ʔigɨrəɲa/*	pedestrian	እግር	እግር
	ተራራማ */tərarama/*	mountainous	ተራራ	ተራራ
	ዛፎቹን */zafotʃuni/*	the trees [*acc*]	ዛፍ	ዛፍ
	ስለዚህም */sɨləzihɨmi/*	therefore [*foc*]	እዚህ	እዚህ

The stem-based and root-based corpora annotations were performed using stem-based and root-based lexicons, respectively. The lexicons were annotated manually whereas documents in the corpus were annotated automatically. Each lexicon contains surface words and their respective annotation forms. Each word in each document was replaced by the corresponding morphologically annotated form using Algorithm 1.

Algorithm 1 Procedure for morphological annotation.

```
Input: stem-based lexicon and row text corpus
Output: stem-based annotated Corpus
Open a lexicon and read line by line
For i→1 to length of document in the corpus:
    Open document D[i] and read words in the document
    For j→ 1 to number of words in document D[i]:
        For k 1 to number of lines in the lexicon:
            Segment line[k] into two using comma
            If a word(j,Di)==Lexicon_word [0][k]:
                Word(j,Di)= Lexicon_word[1][k]
            End if
        End for
    End for
End for
```

5.2. Implementation and Measures

Various experiments were carried out after creating morphologically annotated Amharic text corpora which could be used as benchmarks for developing and evaluating the performance of IR and NLP systems. Python was used for the preprocessing tasks, while indexing and retrieving were performed using Lemur toolkit. The retrieval effectiveness was evaluated automatically using trec_eval tool. Precision-recall curves, precision (P) at a certain level of cut-off (P@5, P@10, P@15 and P@20) and Mean Average Precision (MAP) were used to measure retrieval effectiveness as it is usually the case in IR. Four word embedding models (CBOW and Skip-gram models for stem-based and root-based text representation) were created offline using word2vec by tuning default parameter values. These models were used for query expansion.

5.3. Results

5.3.1. Stopword Identification and Removal

Identification and removal of stopwords are the two preprocessing tasks in many IR systems. In the past, few researches were conducted for creating Amharic stopwords [31]. The stopword lists constructed in this research are evaluated using Lemur on Amharic IR test collection. The removal of stopwords has significant impact on the retrieval effectiveness and index size. For example, the retrieval effectiveness of stem-based and root-based text representations with and without stopwords are presented in Table 8. As shown in Table 8, removal of stopwords has positive impact both in case of stem-based and root-based retrievals. For the sake of comparison between stem-based and root-based text retrieval, stem-based stopword list is created based on morpheme statistics (frequency, mean, variance, and entropy) considering the stem-based morphologically annotated corpus such as the case of root-based stopword list. In the rest of the experiments, we removed stopwords from stem-based and root-based corpora using our stem-based and root-based stopword lists.

Table 8. 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

The overall impact of stopwords and the comparison between stem-based and root-based retrievals are indicated in Figure 2. The top two curves in red and blue represent the root-based and stem-based retrieval, respectively after stopword removal while the bottom two curves (in green and yellow) represent root-based and stem-based retrieval, respectively, with stopwords. Test results have shown that stopword removal improves the effectiveness of Amharic IR on stem-based and root-based text representations. Root-based text representation with stopword removal yields the best results.

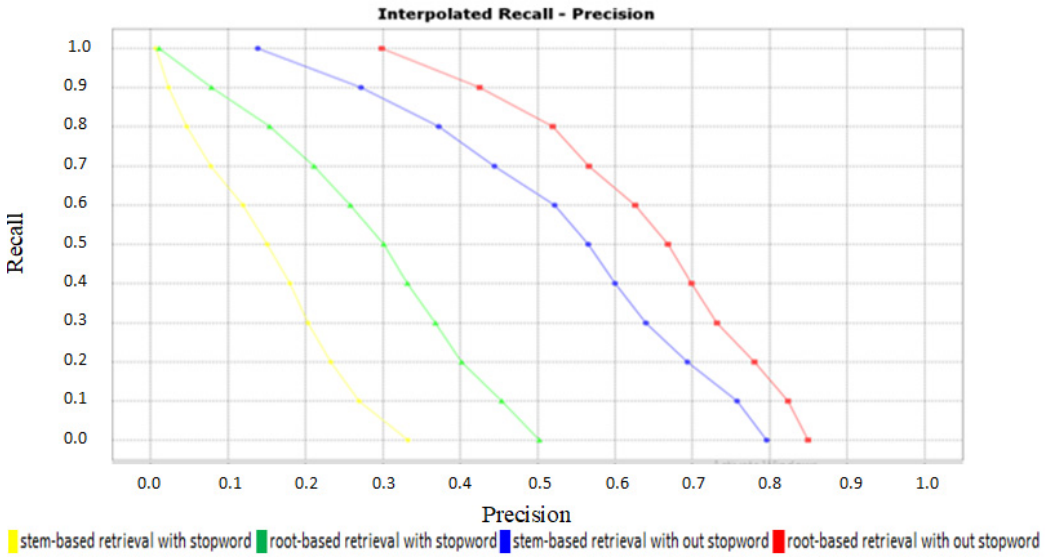


Figure 2. Retrieval effectiveness with and without stopwords.

5.3.2. Test Retrieval Results of Word-Based, Stem-Based and Root-Based Retrieval

Three experiments were conducted for word-based, stem-based, and root-based retrievals. The aim of the experiments was to investigate the effectiveness of morphological analysis (i.e., stem vs. root) on Amharic language. Table 9 reports the effects of morphological analysis on Amharic retrieval effectiveness.

Table 9. Word-based, stem-based, root-based retrieval.

Text Representation	Precision					NDCG
	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

We can observe that morphological analysis has an impact on retrieval effectiveness, both stem and root-based retrievals are more effective than word-based retrieval. The root-based retrieval is the best of the three text representations. As expected, the retrieval effectiveness of the three text representations decreases from precision at 5 documents (P@5) to precision at 20 documents (P@20) also due to the scarcity of relevant documents in the test collection. Root-based retrieval achieves 0.70 for P@20 while stem-based retrieval is 0.53 and word-based retrieval is 0.56.

Figure 3 depicts the overall recall-precision (R-P) curves for the stem-based and root-based text representations. The top line (in blue) is the root-based retrieval effectiveness, whereas the bottom line (in red) is the stem-based one. In this figure, we could see that the performance of the root-based representation is better than stem-based representation at any level of recall.

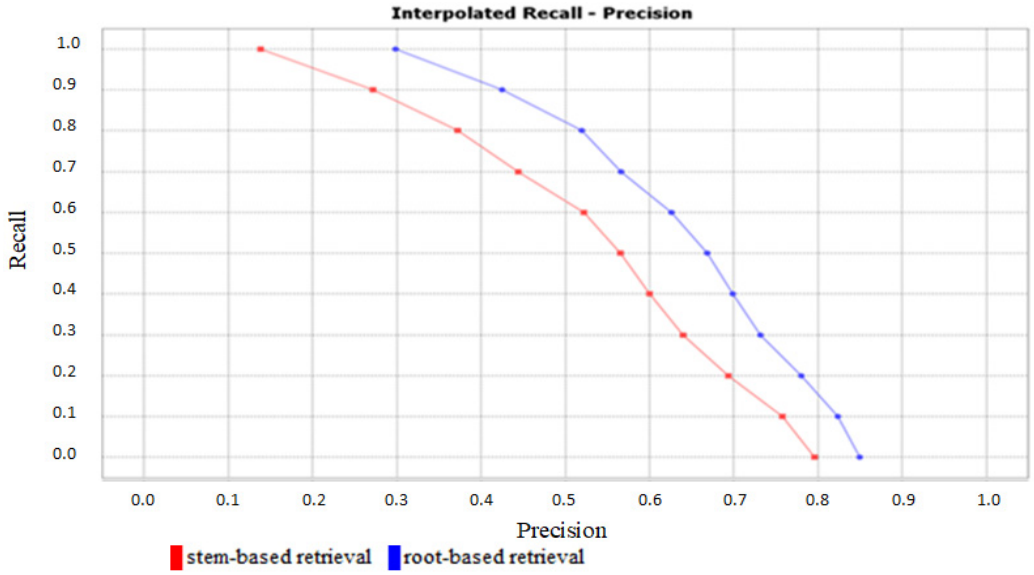


Figure 3. Recall-precision curves for the stem-based and root-based text representations.

5.3.3. Query Expansion Using Word Embedding

CBOW and Skip-gram learning algorithms were trained by tuning the default 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). The retrieval effectiveness of the proposed Amharic IR system before and after query expansion are presented in Table 10.

Table 10. Amharic IR before and after query expansion.

Retrieval Type	Technique	Retrieval Effectiveness				
		P@5	P@10	MAP	R-P	NDCG
Retrieval before query expansion	Stem-based	0.62	0.53	0.57	0.4935	0.71
	Root-based	0.79	0.70	0.70	0.5826	0.86
Retrieval after query expansion	Stem-based CBOW	0.3929	0.3482	0.3159	0.3162	0.5450
	Stem-based Skip-gram	0.4018	0.3597	0.3176	0.3175	0.5498
	Root-based CBOW	0.4507	0.3973	0.4052	0.3842	0.6650
	Root-based skip-gram	0.4434	0.3833	0.3983	0.3608	0.6564

5.4. Discussion

5.4.1. Comparison of Stem-Based and Root-Based Text Representations

When considering average results as depicted in the previous section, the root-based text representation retrieves more relevant documents than the stem-based and word-based text representations. This is certainly because the root-based text representation filters out non-relevant documents more accurately than stem-based and word-based. For example, consider the following queries:

1. የአየር ንብረት ብክለት /*jəʔəjər nibirəti bikiləti* 'air pollution'/
2. የማዳበሪያ ስርጭት /*jəmadabərija sirit∫'iti* 'distribution of fertilizer'/
3. የኤድስ በሽታ የምርመራ እና የምክር አግልግሎት /*jəʔedisi bə∫ita jəmiriməra ʔina jəmikiri ʔəgiligiloti* 'AIDS disease examination and consultation service'/
4. የደን ጭፍጨፋ /*jədəni t∫'if it∫'əfa* 'deforestation'/
5. አልሻባብ ሽብር ጥቃት /*ʔəli∫abab ∫ibir t'ik'*at 'Alshebab terrorist attack'/

The aforementioned queries have retrieval precision as shown in Table 11. Test results indicate that precisions at 5 and 10 are higher for root-based representation than for stem-based.

Table 11. Precision for stem-based and root-based representation.

Query	Stem-Based		Root-Based	
	P@5	P@10	P@5	P@10
1	0.00	0.30	1.00	1.00
2	0.60	0.60	1.00	1.00
3	0.20	0.30	0.60	0.60
4	1.00	0.50	1.00	0.80
5	0.20	0.20	1.00	0.80

The word-based and stem-based methods miss more relevant documents since they could not handle some morphological variations. For example, searching using the stem ሰበር- /*səbəri-* 'break'/ misses documents containing the stem ሰባር- /*səbari-* 'break'/, ስበር- /*sibəri-* 'break'/, and ሱበር- /*səbri-* 'break'/. Despite the stem-based representation 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 relevant retrieved documents. For example, the frequency of the three stems and roots of variants in a given document are presented in Table 12. We can see that the frequency of root is higher than the frequency of stem.

Table 12. Frequency of stem and root in a sample document.

Concept	Frequency		
	Stem	Root	Actual
Action	4	10	10
Dignity	3	7	7
Belief	5	11	11

When looking at the results from which Figure 3 is derived, we found that some relevant documents are retrieved using the root-based representation but not by the stem-based one. Some non-relevant documents are retrieved by the stem-based representation but not by the root-based one. For example, for the query 'የአየር ንብረት ብክለት /*jəʔəjəri nibirəti bikiləti* 'air pollution'/' the root-based retrieval returns 10 relevant documents in the top 10, while the stem-based text representation returns only 3 relevant documents in the top 10. The reasons are the three strong sides of root over stem as described in what follows.

i. Root can conflate all morphologically variants to one common root. For example, the stems of variants ሰበረ /*səbərə*/, ስበር /*sibəri*/, ስብራት /*sibirati*/, አሳበረ /*ʔəsabərə*/, ሰባራ

/*səbara*/, አሳቢራት /*ʔəsabirati*/, ተሳቢሮ /*təsəbiro*/ are ሰበር- /*səbəri-*/, ስበር- /*sibəri-*/, ስብር- /*sibiri-*/, ሳብር- /*sabari-*/, ሰባር- /*səbari-*/, ሳብር- /*sabiri-*/, and ሰበር /*səbəri-*/, respectively. This creates vocabulary mismatch. Many words have similar cases in Amharic. 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 and rank.

ii. Root does not conflate semantically unrelated words to a common form. However, this is not the case for the stem-based representation which may conflate semantically unrelated words to a common form. For example, ገደል /*gədəli*/ is the stem of the verb ገደለ /*gədələ* 'he killed'/ and the noun ገደሎች /*gədəlotʃi* 'cracks'/. The verb ሲገድል /*sigədili* 'as he kills'/ and the noun ገድል /*gədili* 'contending'/ have the same stem ገድል- /*gədili-*/. This leads to retrieval of non-relevant documents when using stems, which decreases precision.

iii. The retrieval results of the stem-based representation depend on the query word variants, while the root-based representation does not. The root-based representation performs equally for all the variants of the query terms. However, the stem-based representation returns different results in different ranks. Users would certainly construct the same information need using different queries, using different word variants. However, in Amharic, query terms of word variants might not be similar after stemming. For example, a query 'the causes of air pollution' could be constructed as 'የአየር ንብረት ብክለት መንስኤዎች /*jəʔjəri nibirəti bikiləti mənisiʔewotʃi*/' or 'የአየር ንብረት በካይ መንስኤዎች /*jəʔjəri nibirəti bəkaji mənisiʔewotʃi*/'. In stem-based representation, the two queries have {አየር, ንብረት, ብክል, መንስኤ} and {አየር, ንብረት, በካል, መንስኤ} query terms, respectively, in which case the system retrieves different results. In the case of root-based representation, the two queries have the same set of query terms {አየር, ንብረት, ብ-ክ-ል, መንስኤ}, which guarantees the same retrieval result.

Several queries may be formulated by users with variants of words. To investigate the effect of stems of variants of words in queries on the overall Amharic IR, we created an alternative form of each query in our test collection and made the comparison between the two options. The two stem-based forms represent variants of words in users' queries. All possible options of a query formulated from variants of words have the same root-based representation though they might have more than one stem representation. As depicted in Figure 4, the two query options in stem-based representation performed differently on the same test collection, but performed equally in case of root-based representation. The top curve (in green) is the root-based retrieval whereas the remaining two are the stem-based retrievals. The stem-based retrieval curves might be even changed if our queries in the topic set are constructed just by changing only variants of one or more words in each query.

5.4.2. Comparison of Amharic IR before and after Query Expansion

Stem-based and root-based retrieval effectiveness before and after query expansion are investigated on the same corpus using the same retrieval model that is LM. Figure 5 depicted the overall retrieval effectiveness of Amharic IR system with and without query expansion on the same test collection and retrieval mode. The red, blue, aqua, purple, yellow and green colors represent root-based retrieval without query expansion, stem-based retrieval without query expansion, root-based retrieval after query expansion using Skip-gram, root-based retrieval after query expansion using CBOW, stem-based retrieval after query expansion using CBOW, and stem-based retrieval after query expansion using Skip-gram algorithm. As shown in Table 10 and Figure 5, Amharic IR retrieval after query expansion is less than conventional retrieval. Some expanded terms of an initial query term are related syntactically rather than semantically. These syntactical related terms affect the retrieval effectiveness of Amharic IR. For example, the city name ባህርዳር /*bahiridari*/ is expanded with other city names መቀሌ /*məkʼəle*/, ሀዋሳ /*hawasa*/, ጎንደር /*gonidəri*/, ጎርጎራ /*gorigora*/ and ነቀምት /*nəkʼəmiti*/. Due to this the system retrieves those documents discussing about these cities though they are irrelevant to the user need. Many cases such

as these exist in the corpus which leads to performance decrement. The other reason for low performance is ambiguity of some terms. For example, the stem of the words ስማቸው /*sɨmatʃəwɨ* 'their name' or 'she kissed them'/ is ስም /*sɨmɨ'*/; the stem of the word ፈረሱ /*fərəsu* 'the horse' or 'they are destroyed'/ is ፈረስ /*fərəsɨ*/; the stem of the word አፈሩ /*ʔəfəru* 'the soil' or 'they are ashamed'/. Many cases such as these occur in the language. This has shown negative impact on retrieval result. As shown in figure, CBOW and Skip-gram algorithms perform closely except CBOW performs better slightly on stem-based text representation.

Even though root-based query expansion is lower than conventional retrieval, it is slightly better than stem-based retrieval. The reason is that the probability of the root of a word appearing with the root of a related word is greater than or equal to the stems of related words.

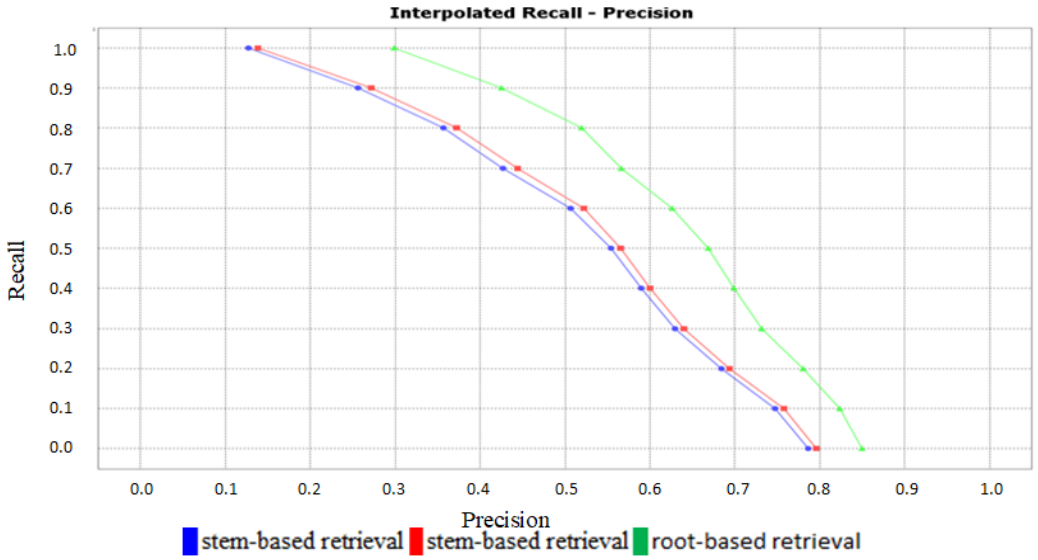


Figure 4. Recall-precision curves for the stem-based and root-based text representations.

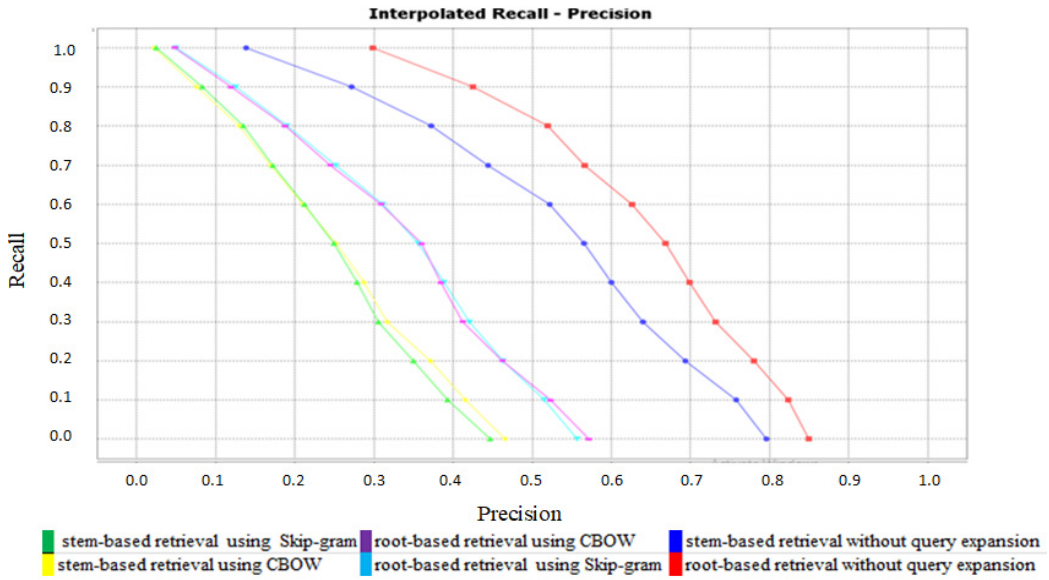


Figure 5. Recall-precision curves before and after query expansion using word embedding.

5.4.3. Comparison with Previous Amharic IR Studies

Few Amharic IR systems were developed in the past. The majority of them are based on stems [26,50,55]; some others are based on citation form [46] and roots [25]. However,

due to the complexity of the language, the stem-based, and citation form, and n-gram models do not work well.

In our experiments, we found that roots are more powerful for Amharic IR than stems. Other authors suggested stem-based as the best option. Alemayehu and Willett [25] investigated the effects of the stem and root-based text representations on Amharic IR. The finding was that the stem-based is better than root-based representation. The justification was that many Amharic words have common root though they are semantically unrelated. The report states that stem-based improves retrieval effectiveness in terms of rejecting irrelevant documents that could be retrieved by the root-based method. As a result, [25] reject root-based representation. However, their research has some limitations; the main issue being the way they represented roots. For example, for the root representation of the word ዝናብ /*zinabi* 'rain'/ they used ዝንብ (which should have been ዝ-ን-ብ /*z-n-b*/), producing a similar term as ዝንብ /*zinibi* 'fly'/. Another issue is the use of same root representation for different words leading to conflation of many semantically unrelated words. In their work, all word classes are represented using radicals, which should have been the case only for words derived from verbal roots. Furthermore, weak consonants were ignored during the extraction of roots producing incorrect representation for some words. For example, ምት /*mti*/ was considered as the root form of both ሞተ /*motə* 'he died'/ and መታ /*məta* 'he hit'/, which should have been ም-ው-ት /*m-w-t*/ and ም-ት /*m-t*/, respectively. Such inaccurate text representations would significantly affect the retrieval performance. Previously, the impact of query expansion using word embedding was evaluated and obtained 0.23 MAP [50]. We obtained better precision both in case of CBOW and Skip-gram models on stem-based and root-based corpora (see Table 10). One of the main reason is we handle the unique morphological characteristics of Amharic language and following slightly different Amharic IR system design from conventional IR system design. Moreover, we used LM and run more queries on the test collection.

5.4.4. Comparison with Google Amharic Search Engine

The Google Amharic search engine is based on stems. It returns different retrieval results in different ranks for the same query using different variants of the query words. This is the same for our stem-based representation. For example, Google search results for the queries የአጥንት ስብራት /*jəʔtʼinɨti sɨbirati* 'the state of being broken/ and የአጥንት መሰበር /*jəʔtʼinɨti məsəbəri* 'fracturing of bone'/ are different though the same concept is expressed via different variants. The top 4 search results for both queries are shown in Table 13. Our text representations differ from Google search engine into two ways:

i. Google is based on both basic stems and derived stems. It returns different results for basic stem and derived stem queries though they are semantically similar. However, in our research the stem-based representation is based on basic stems only; we thus acquire the same retrieval results for both basic stems and derived stems. For example, retrieving using the basic ሰበር- /*səbəri-* 'break'/ returns documents containing derived stems such as መሰበር- /*məsəbəri-* 'to break'/, ሰባበር- /*səbabəri-* 'break repeatedly'/, ተሰበር- /*təsəbəri-* 'is broken'/, and አሰበር- /*ʔəsəbəri-* 'he is the cause of breaking'/.

ii. Google does not employ roots to represent verbs and words derived from them. However, root is the best word form to conflate all variants.

Table 13. The top 4 search results from Google Amharic for two queries.

Query	የአጥንት ስብራት /'the state of bone being fractured'/	የአጥንት መሰበር /'fracturing of bone'/
Retrieval Result	የአጥንት ስብራት-YouTube የአጥንት ስብራት-Amharic search engine የአጥንት ስብራት የእጅ የአጥንት ስብራትና መዉሰድ ያለበት ጥንቃቄ	የአጥንት ስብራትና መዉሰድ ያለበት ጥንቃቄ የአጥንት መሳሰት ህመም አስቲዮ ፓሮሲስ ደምጽ አልባው በሽታ- የመጠበቂያ አስቲዮ ፓሮሲስ ደምጽ አልባው በሽታ- የኢትዮጵያ

5.4.5. Comparison of LM and BM25

We compared the performance of language modeling (LM) and BM25 using the root-based text representation. Although LM is considered as a strong baseline for IR in different languages, previous Amharic IR research was based on vector space [25,26,55] but not on LM. We used KL-divergence (LM) with default parameters (i.e., Dirichlet priority smoothing method (1000) and Interpolate smoothing strategies μ = 2000) and BM25 with the following parameters values: BM25 k1 = 1.2; BM25 B = 0.75; BM25 K3 = 1000; TF weighting = 0.5 and number of feedback = 50. As depicted in Figure 6, the LM line graph (in red) is above the BM25 (in blue) graph. Both precision and recall values of LM are higher than BM25 at different levels. This is certainly because of the capability of LM to capture words dependency and estimate the probability distribution of a query in each document. LM seems to be better for Amharic language retrieval.

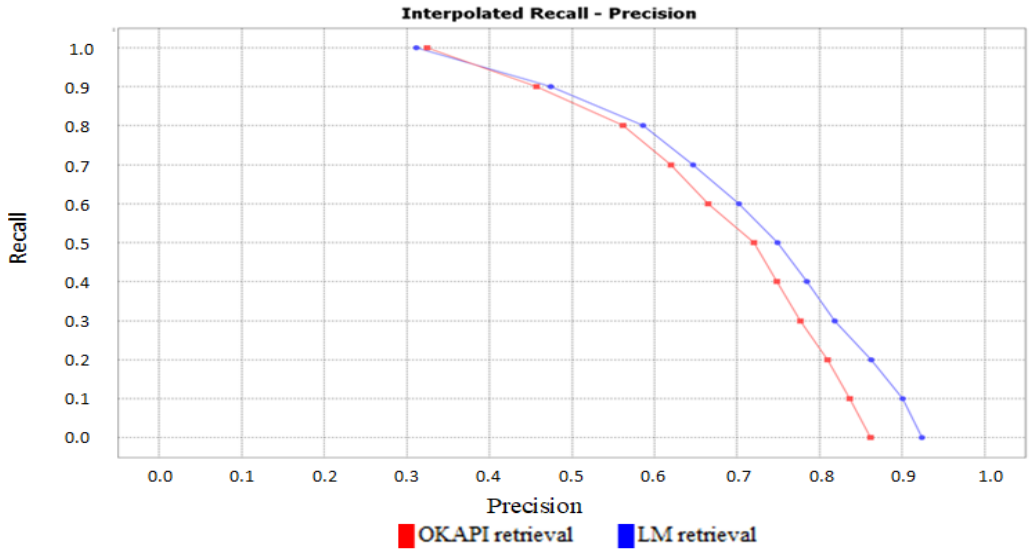


Figure 6. The performance of LM and BM25.

5.4.6. Comparison with Other Semitic Languages

Semitic languages have a complex morphology. In many retrieval systems developed for Semitic languages, the common trend to analyze texts is to make use of stemming. For example, a comparison between the effects of stem and root on Arabic IR was carried on in [8,13,37]. Even though the stem-based representation misses some relevant documents, it was found to be better than root-based representation. It was reported that the root-based representation retrieves many non-relevant documents. In Amharic stem-based representation misses many relevant documents and retrieves non-relevant documents. The root-based representation increases the retrieval of relevant documents and minimizes the retrieval of non-relevant documents to a user query.

The retrieval performance of the proposed system may be further improved by using NLP tools such as co-reference resolution, anaphora resolution, word sense disambiguation and spelling checker. Since these Amharic NLP tools are unavailable in usable form, we were unable to integrate them into our system. This limitation might have affected the retrieval performance. On the other hand, the result of query expansion using neural word embedding can be improved by increasing the dataset.

6. Conclusions

Amharic is a morphologically complex and under-resourced language. This poses tremendous challenges for natural language processing and information retrieval. The complexity of Amharic makes it difficult to transfer existing NLP tools developed for other languages to Amharic. Although Amharic is a working language for more than

117 million people, the scarcity of standard IR corpora and resources limited research and development works in processing Amharic text. In this paper, we present a study on plausible text representation for Amharic IR. We performed morphological analysis for handling variation while indexing and retrieving text documents in an *adhoc* task. It is shown that the morphology of the language affects the retrieval effectiveness of Amharic IR. Through experiments, we found that the root-based text representation is better than the stem-based one. Root-based is shown to be robust for conflating variants and retrieving more relevant documents. Another major contribution of our research is the construction of corpora and resources for IR and related researches that could be used as a standard reference. We built an Amharic IR test collection, stopword list and morphologically annotated corpora, which are valuable and important to foster research in Amharic IR. Future work is directed at the use of word sense disambiguation to resolve term ambiguities that occur during query expansion.

Author Contributions: Conceptualization, T.Y., J.M. and Y.A.; formal analysis, T.Y. and Y.A.; funding acquisition, J.M.; investigation, T.Y.; methodology, J.M.; resources, T.Y.; software, T.Y.; supervision, J.M. and Y.A.; validation, T.Y., J.M. and Y.A.; writing—original draft, T.Y.; writing—review and editing, J.M. and Y.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Experiment dataset and resources are available online at https://www.irit.fr/AmharicResources/ accessed on 22 December 2021.

Conflicts of Interest: The authors declare no conflict of interest.

Notations

The following notations are used in this article.

1	first person
2	second person
3	third person
s	singular
p	plural
pre	preposition
foc	focus
nom	nominative
conj	conjunction
neg	negative
gen	genitive
def	definite marker
adj	adjectivizer
ncmp	negative complement
pos	possessive
m	masculine
pas	passive
pal	palatalizater
inf	infinitive
cau	causative
com	complement

Appendix A. የአጥንት ስብራትና መወሰድ ያለበት ጥንቃቄ

በለትተለት እንቅስቃሴያችን ውስጥ በድንገትና ባላሰብነው ሁኔታ ከመለስተኛ አደጋዎች እስከ ከባድ የአጥንት መሰባበር ሊደርስብን ይችላል። የስብራት አደጋ ዋናው ችግሩ የሚያስከትለው ህመምና ስቃይ ከፍተኛ መሆኑ ነው። በተለይ ጉዳቱ በገጠመን ቦታ እና አካባቢው ላይ የሚገኙት የነርቭ ጫፎች፣ ትንሽ እንኳን ሲነኩ ከፍተኛ ህመም ይኖራቸዋል። በእንዲህ አይነት ችግር ውስጥ ያለ ሰው እንቅስቃሴን በፍፁም ማስወገድ አለበት። ሌላው የዚህ አደጋ ትልቁ ችግር፣ የሰውነት ውስጥ ለውስጥ ደም መፍሰስ ሊያስከትል መቻሉ ነው። ይህ ሁኔታ ከባድ በመሆኑ እራስን ሊያስትና የከፋ የጤና ችግር ሊያስከትል ይችላል። የአጥንት መሰበር በፍጥነት ህክምና አግኝቶ እንክብካቤ ካልተደረገለት ወደ ተወሳሰበ ሁኔታ ሊያመራ እንደችል ማወቅ ያስፈልጋል። ይሁን እንጅ ሰባራ አጥንት አስፈላጊውን እንክብካቤና ምርመራ ከተደረገለት ቶሎ ይድናል። በስብራት ምክኒያት የአከባቢያዊ ሕብረ ሕዋሳት ጉዳት ከደረሰባቸው ከፍተኛ ችግር ያሰክትላል። የአጥንት መሰበር ብዙ ደም መፍሰስ እና መሰብሰብ ለፈጥር ይችላል። ድንገተኛ ስብራት ሲያጋጥመን ቶሎ ህክምና መውሰድና ከከፋ ጉዳት መጠንቀቅ አለብን። ሳይንሳዊ ያልሆነ ህክምና ጤነኛን አጥንት እንደመስበር ይቆጠራል። ስለዚህ ሀኪምን ማማከር ይመከራል። በህክምና ባለሙያ መመርመር ያስፈልጋል። የአጥንት መሰበር በሁለት መንገድ ሊያጋጥመን ይችላል። አንደኛ፣ የቆዳ እና ሥጋ አካል ክፍላችንን ዘልቆ በመግባት የሚደርስ የአጥንት መሰበርና የቆዳ ማበጥ ሲሆን፣ በዚህም ቆዳችን የመቁሰል ችግር ያጋጥመዋል። የቆሰለ ቆዳ ቆስሎ ስለማይቀር ሊመፈንዳ ይችላል። በዚህ ጉዳት አጥንታችን፣ ስጋችንን ቀዶ ፈጦ ሊወጣ ይችላል። በእንዲህ ሁኔታ የሚፈጠር ቁስለት ኢንፌክሽን የመፍጠር እና ብዙ የሰውነት ክፍል ሊያዳርስ ስለሚችል የመዳን ሂደትን ሊያረዝም እና ሊያወሳስብ ይችላል። ስለሆነም ቶሎ እርዳታ ለማግኘት እንሞክር። ኢንፌክሽኑን ለመከላከል ቁስሉን በፀረ-አንቲሴፕቲክ ፈዋሽ መድኀኒት መጠቀም አለብን። ከዚህ በተጨማሪ የደም ሥሮችን የሚያጠናክሩ መድሀኒቶችን መጠቀምና ቆሳይ ቆዳን መንከባከብ የኖርብናል። ስብራትን ለመደገፍ የምንጠቀማቸው የተለያዩ ቁሶች ሰውነት እያበጠ ሲመጣ ህመም ሊፈጥሩ ይችላሉ። ሰባራ አጥንት እንዳይነቃነቅ ሸምበቆ ተጠቅሞ መደገፍ አለበት። ድንገተኛ ስብራት የገጠመው ምንም እንቅስቃሴ እንዳያደርግ ተደርጎ በተለያዩ ቁሶች መታሰር አለበት። ለባስ ችግር እንዳንዳረግና ስብራት እንዳይመለስ ተደርጎ መታሰር አለበት። አንድ ተጎጂ እግሩንና እጁን ማንቀሳቀስ ካልቻለና ሲነኩት ስሜት አልባ ከሆነ የገጠመው የአከርካሪ አጥንት ስብራት ሊሆን ይችላል። አካሉን ስንነካው ምንም ካልተሰማው እና ምንም ምላሽ የማይሰጥ ከሆነ ሁኔታው ከባድ ስለሆነ ምንም እንዳይንቀሳቀስ አድርገን የሕክምና እርዳታ ማግኘት እስከሚችል ድረስ እንዲያርፍ ማድረግ አለብን። ጉዳተኛውን ወደ ተሻለ ቦታ መውሰድ ወይም የሕክምና እርዳታ ሊያገኝ ወደሚችልበት ቦታ ማጓጓዝ ካስፈለገ፣ ከፍተኛ ጥንቃቄ መደረግ አለበት፤ እንዲህ ዓይነት ጉዳት የገጠመውን ሰው ስናጓጉዘው ምንም ሊያንቀሳቅሰው በማይችል አልጋ ወይም ቃሬዛ ላይ ጭነነው መሆን አለበት። ችግሩ በእግር ላይ ከሆነ፣ ሙሉ የእግር መዋቅር ምንም እንቅስቃሴ እንዳያደርግ ተደርጎ በተለያዩ ቁሶች መታሰር አለበት። ምንም ዓይነት ቁስ በእጃችን ባይገኝ እንኳን ስብራት ካልገጠመው ጤነኛ እግር ጋር ድጋፍ እንዲሆነው ተደርጎ ተጣምሮ መታሰር አለበት። ስብራቱ ከገጠመው አካል በላይ እና በታች እንዲሁም መጋጠሚያ ጋር ንቅናቄን እንዳይፈጥር አድርገን ማሰር አለብን። ሁኔታውን በየጊዜው በመከታተል የማስተካከያ እርምጃ ማድረግ ተገቢ ነው። ቶለ ከታከመ ስብራቱ ቀስ በቀስ ይገጥማል ። ለዚህ ችግር መጀመሪያ መደረግ ያለበት የተሰበረውን አጥንት ወደ ቦታው በመመለስ እንደ ሸምበቆ ወይም እንጨት ያሉ ቁሶችን ተጠቅሞ እንዳይንቀሳቀስ ወይም እንዳይነቃነቁ አድርጎ ማሰር ነው። ቁስለቱን እንደ ማንኛውም ቁስል በማፅዳት እና በንፁህ ጨርቅ በመጠቅለል ወደ ሕክምና እስከምንደርስ ድረስ መከታተል ነው። በዚህ መልኩ እርዳታ ያገኘ ተጎጂ እራሱን ከእንቅስቃሴ በመግታት ማገገም መቻል አለበት፤ አለበለዚያ ሁኔታው ሊባባስ ይችላል። ሁለተኛው ደግሞ የሰውነታችንን የሥጋ ክፍል ዘልቆ ሳይገባ ወይም ቁስለት ሳይፈጥር የሚገጥመን የማይታይ የአጥንት መሰበር ወይም መሰንጠቅ ነው። ሁኔታውን ለመረዳት ስዕሎቹን ተመልከቱ፤ በጉዳት ወቅት የሚከተሉት ምልክቶች የአጥንት መሰበርን ወይም መሰንጠቅን ሊጠቁመን ይችላሉ። አደጋው ባጋጠመን አካል ወይም ቦታ ላይ ፡- ከፍተኛ ህመም, በቆዳችን ላይ የሚፈጠር የቀለም ለውጥ, እብጠት, ቅርፅ ለውጥ። ሌላው ስብራት ሊገጥመን የሚችልበት ስብራቶች በሚታከሙበት ወይም ወደቀድሞ ይዘታቸው ለመመለስ በሚሞከርበት ጊዜ ነው። ስብራቶች በሚታከሙበት ወይም ወደቀድሞ ይዘታቸው ለመመለስ በሚሞከርበት ጊዜ የሕመም ስቃያቸው ከባድ ስለሚሆን፣ ተጎጂው ራሱን ስቶ የሚገኝ ከሆነ ሳይነቃ በፍጥነት ተስተካክለው መልክ ቢይዙ የሚመረጥ ነው። ስብራትን ለመደገፍ የምንጠቀማቸው የተለያዩ ቁሶች ሰውነት እያበጠ ሲመጣ ህመም ሊፈጥሩ ስለሚችሉ፣ ለስላሳ የሆኑ ነገሮች በገላና በመደገፊያዎቹ መካከል ብንጎዘጉዝ ሕመሙን ይቀንሰዋል። እንዲሁም የውጥረት መጠኑን ክትትል በማድረግ የደም ዝውውርን እንዳይገታ ማላላት ያስፈልጋል። ስብራት ተስተካክሎ አንዴ ከታሰረም በኃላ እንዳይንቀሳቀስ ተደርጎ በደንብ መታሰር አለበት፤ ነገር ግን ምንጊዜም የደም ዝውውር ላይ ዕክል እንዳይፈጥር በመከታተል የማላላት እርምት መወሰድ መዘንጋት የለብንም።

References

1. Sanderson, M.; Croft, W. The history of information retrieval research. *Proc. IEEE* **2012**, *100*, 1444–1451. [CrossRef]
2. Buckley, C.; Voorhees, E. Retrieval system evaluation. In *TREC: Experiment and Evaluation in Information Retrieval*; MIT Press: Cambridge, MA, USA, 2005; Volume 32, pp. 53–75, Chapter 3.
3. Ferro, N. CLEF 15th Birthday: Past, Present, and Future. *ACM SIGIR Forum* **2014**, *48*, 31–55. [CrossRef]
4. Clough, P.; Sanderson, M. Evaluating the performance of information retrieval systems using test collections. *Inf. Res.* **2013**, *18*, 1–13.
5. Cleverdon, C.W. The evaluation of systems used in information retrieval. In *Proceeding of the International Conference on Scientific Information*; The National Academies Press: Washington, DC, USA, 1959; pp. 687–698.
6. Harman, D. Overview of the second text retrieval conference (TREC-2). *Inf. Processing Manag.* **1995**, *31*, 271–289. [CrossRef]
7. Kando, N.; Kuriyama, K.; Nozue, T.; Eguchi, K.; Kato, H.; Adachi, J. The NTCIR Workshop: The first Evaluation Workshop on Japanese Text Retrieval and Cross-Lingual Information Retrieval. In Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages, Tokyo, Japan, 15 December 1999; pp. INV-1–INV-7.
8. 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, QC, Canada, 2002.
9. Smeaton, A. Progress in the application of natural language processing to information retrieval tasks. *Comput. J.* **1992**, *35*, 268–278. [CrossRef]

10. Jackson, P.; Moulinier, I. *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*, 5th ed.; John Benjamins Publishing: Amsterdam, The Netherlands, 2007.
11. Cambria, E.; White, B. Jumping NLP curves: A review of natural language processing research. *IEEE Comput. Intell. Mag.* **2014**, *9*, 48–57. [CrossRef]
12. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
13. 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, Leicester, UK, 2000.
14. El-Haj, M.; Kruschwitz, U.; Fox, C. Creating language resources for under-resourced languages: Methodologies, and experiments with Arabic. *J. Lang. Resour. Eval.* **2015**, *49*, 549–580. [CrossRef]
15. Yifru, M.; Tefera, S.; Besacier, L. Part-of-speech tagging for under-resourced and morphologically rich languages: The case of Amharic, HLTD. In Proceedings of the International Conference on Human Language Technology for Development, Alexandria, Egypt, 2–5 May 2011; pp. 50–55.
16. 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)*; Europeans Language Resources Association: Istanbul, Turkey, 2012; pp. 7–12.
17. Assabie, Y. *Development of Amharic Morphological Analyzer*; Technical Report; Ethiopia Ministry of Communication and Information Technology: Addis Ababa, Ethiopia, 2017.
18. Alemayehu, N.; Willett, P. Stemming of Amharic words for information retrieval. *J. Lit. Linguist. Comput.* **2002**, *17*, 1–17. [CrossRef]
19. 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; Association for Computational Linguistics: Prague, Czech Republic, 2007; pp. 104–110.
20. Amsalu, S.; Gibbon, D. Finite state morphology of Amharic. In Proceedings of the 5th Recent Advances in Natural Language Processing, Borovets, Bulgaria, 21–23 September 2005; pp. 47–51.
21. Gasser, M. Horn Morpho: A system for morphological processing of Amharic, Oromo, and Tigrinya. In Proceedings of the Conference on Human Language Technology for Development, Alexandria, Egypt, 2–5 May 2011; pp. 94–99.
22. Abate, M.; Assabie, Y. Development of Amharic morphological analyzer using memory-based learning. In Proceedings of the 9th International Conference on Natural Language Processing, Warsaw, Poland, 17–19 September 2014; pp. 1–13.
23. Sisay, F. Part-of-speech tagging for Amharic using conditional random fields. In Proceedings of the ACL-2005 Workshop on Computational Approaches to Semitic Languages, Ann Arbor, MI, USA, 29 June 2005; pp. 47–54.
24. Gamback, B.; Olsson, F.; Alemu, A.; Asker, L. Methods for Amharic part-of-speech tagging. In Proceedings of the 1st Workshop on Language Technologies for African Languages; Association for Computational Linguistics: Stroudsburg, PA, USA, 2009; pp. 104–111.
25. Alemayehu, N.; Willett, P. The effectiveness of stemming for information retrieval in Amharic. *Program Electron. Libr. Inf. Syst.* **2003**, *37*, 254–259. [CrossRef]
26. 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, Marrakech, Morocco, 29 November–4 December 2009; pp. 318–325.
27. Samuel, E.; Bjorn, G. Classifying Amharic news text using self-organizing maps. In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, Ann Arbor, MI, USA, 29 June 2005; pp. 71–78.
28. Gamback, B. Tagging and verifying an Amharic news corpus. In Proceedings of the 8th International Conference on Language Technology for Normalization of Less-Resourced Languages, Istanbul, Turkey, 22 May 2012; pp. 79–84.
29. Abate, S.T.; Melese, M.; Tachbelie, M.Y.; Meshesha, M.; Atinafu, S.; Mulugeta, W.; Assabie, Y.; Abera, H.; Seyoum, B.E.; Abebe, T.; et al. Parallel corpora for Bi-Lingual English-Ethiopian languages statistical machine translation. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 3102–3111.
30. Demeke, G.; Getachew, M. Manual annotation of Amharic news items with part-of-speech tags and its challenges. *ELRC Work. Pap.* **2006**, *2*, 1–16.
31. Yeshambel, T.; Mothe, J.; Assabie, Y. Evaluation of corpora, resources and tools for Amharic information retrieval. In Proceedings of the 8th EAI International Conference on Advancements of Science and Technology, Bahir Dar, Ethiopia, 2–4 October 2020; pp. 470–483.
32. Yeshambel, T.; Mothe, J.; Assabie, Y. Amharic document representation for *adhoc* retrieval. In Proceedings of the 12th International Conference on Knowledge Discovery and Information Retrieval, Budapest, Hungary, 2–4 November 2020; pp. 124–134.
33. Khoufi, N.; Boudokhane, M. Statistical-based system for morphological annotation of Arabic texts. In Proceedings of the Student Paper Workshop Associated with RANLP 2013, Hissar, Bulgaria, 9–11 September 2013; pp. 100–106.
34. Countrymeters. Ethiopian Population. Available online: http://countrymeters.info/en/ethiopia (accessed on 20 June 2021).
35. Hetzron, R. Ethiopian Semitic: Studies in classification. *J. Semit. Study Monogr.* **1972**, *7*, 139–141.

36. Yifru, M.; Wolfgang, M. Morphology-Based Language Modeling for Amharic. Ph.D. Dissertation, University of Hamburg, Departments of Informatics, Hamburg, Germany, 2010.
37. Larkey, L.; Ballesteros, L.; Connell, M. Light stemming for Arabic information retrieval. In *Arabic Computational Morphology*; Springer: Dordrecht, The Netherlands, 2007; Volume 38, pp. 221–243.
38. Ambati, V.; Rohini, U.; Pramod, P.; Balakrishnan, N.; Reddy, R. Multilingual information access: Information Retrieval and Translation in a Digital Library. In Proceedings of the 2nd International Conference on Universal Digital Library, Alexandria, Egypt, 8–10 January 2006.
39. Darwish, K.; Magdy, W. Arabic information retrieval. *Found. Trends Inf. Retr.* **2014**, *7*, 239–342. [CrossRef]
40. Abdusalam, A. Effective Retrieval Techniques for Arabic Text. Ph.D. Dissertation, RMIT University, Melbourne, VIC, Australia, 2008.
41. Al-Hadid, I.; Afaneh, S.; Al-Tarawneh, H.; Al-Malahmeh, H. Arabic information retrieval system using the neural network model. *Int. J. Adv. Res. Comput. Commun. Eng.* **2014**, *3*, 8664–8668. [CrossRef]
42. Alnaied, A.; Elbendak, M.; Bulbul, A. An intelligent use of stemmer and morphology analysis for Arabic information retrieval. *Egypt. Inform. J.* **2020**, *21*, 209–217. [CrossRef]
43. Carmel, D.; Maarek, Y. Morphological disambiguation for Hebrew search systems. In *Proceedings of the 4th International Workshop on Next Generation Information Technologies and Systems (NGITS 1999)*; Springer: Berlin/Heidelberg, Germany, 1999; pp. 312–325.
44. Ornan, U. A morphological, syntactic and semantic search engine for Hebrew texts. In Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages, Philadelphia, PA, USA, 11 July 2002; pp. 1–10.
45. Argaw, A.; Asker, L.; Cöster, R.; Karlgren, J.; Sahlgren, M. *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; Springer: Berlin/Heidelberg, Germany, 2006; pp. 83–92.
46. Argaw, A.; Asker, L. Amharic-English information retrieval. In *Workshop of the Cross-Language Evaluation Forum for European Languages*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 143–149.
47. Chifu, G.; Hristea, F.; Mothe, J.; Popescu, M. Word sense discrimination in information retrieval: A spectral clustering-based approach. *Inf. Processing Manag.* **2015**, *51*, 16–31. [CrossRef]
48. Aklouche, B.; Bounhas, I.; Slimani, Y. Query expansion based on NLP and word embedings. In Proceedings of the 27th Text REtrieval Conference, TREC 2018, Gaithersburg, MD, USA, 14–16 November 2018.
49. Diaz, F.; Bhaskar, M.; Nick, C. Query expansion with locally-trained word embeddings. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; pp. 367–377.
50. Getnet, B.; Assabie, Y. Amharic information retrieval based on query expansion using semantic vocabulary. In Proceedings of the 8th EAI International Conference on Advancements of Science and Technology, Bahir Dar, Ethiopia, 2–4 October 2021; pp. 407–416.
51. Hadi, Y.; Noa, S.; Mohd, M.; Atwan, J. Word-embedding-based query expansion: Incorporating deep averaging networks in Arabic document retrieval. *J. Inf. Sci.* **2021**, *46*, 1–19. [CrossRef]
52. Balakrishnan, V.; Lloyd-Yemoh, E. Stemming and lemmatization: A comparison of retrieval performances. *Lect. Notes Softw. Eng.* **2014**, *2*, 262–267. [CrossRef]
53. Yimam, B. *Yamarigna Sewasiw (Amharic Grammar)*, 2nd ed.; CASE: Addis Ababa, Ethiopia, 2000.
54. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; Advances in Neural Information Processing Systems. Volume 2, pp. 3111–3119.
55. 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, 28–31 October 2012; pp. 32–39.