Amharic Information Retrieval Based on Query Expansion Using Semantic Vocabulary

Berihun Getnet[1(✉)] and Yaregal Assabie[2]

[1] Department of Computer Science, Wolkite University, Wolkite, Ethiopia
berihun.getnet@wku.edu.et
[2] Department of Computer Science, Addis Ababa University, Addis Ababa, Ethiopia
yaregal.assabie@aau.edu.et

Abstract. The increase in large scale data available from different sources has demanded advancement in information retrieval. As a result, information retrieval based on learning from high dimensional vectors based on the words adjacent to other words or surrounding terms has become more attractive in recent times. The meaning is extracted from the context of words but not using the actual sense of words. The system responds to relevant results for the users by expanding the original queries from semantic lexical resources constructed automatically from a text corpus using neural word embedding. In this study, we propose query expansion for Amharic information retrieval using semantic vocabulary. The semantic vocabulary is automatically constructed from a text corpus using neural word embedding. The user's query is expanded based on the word analog prediction. Information retrieval using semantic vocabulary based on ranked and unranked retrieval increases by a recall of 24 and 15%, respectively albeit at the expense of some precision.

Keywords: Query expansion · Semantic vocabulary · Word embedding · Amharic information retrieval

1 Introduction

Due to a growing need of users to search for relevant documents, much attention has been given over the years to improve the performance of information retrieval systems. Query expansion has been one of the core issues considered for improvement of information retrieval systems. It is the process of reformulating the original query into a set of additional multiple similar terms to retrieve more relevant documents from data sources [1]. Query expansion is traditionally implemented by making use of manually constructed linguistic resources such as thesaurus, dictionary and WordNet [2]. Query expansion based on semantics has recently become more popular due to the understanding that meaning is subjective [1, 3]. However, manual construction of such resources is difficult as the task is labor-intensive and time consuming. On the other hand, the availability of large corpora and high-speed processing power of computers has brought an opportunity

M. A. Delele et al. (Eds.): ICAST 2020, LNICST 384, pp. 407–416, 2021.
https://doi.org/10.1007/978-3-030-80621-7_29

for automatic construction of linguistic resources. Thus, there has been a growing interest among researchers and developers in automation and data-driven linguistic analysis where lexical resources like semantic vocabulary, thesaurus, dictionary, and WordNet can be constructed either automatically from natural language texts [3–5].

One of the theories effectively applied for analyzing linguistic data is distributional semantics. Distributional semantics assumes that the meaning is extracted from word contexts and its distributions across a vector space [6–8]. In this case, the meaning of words is extracted from the text considering that words occurring in similar contexts are semantically similar. Thus, semantic vocabulary can be constructed automatically from a corpus based on the assumption that words distributed across a multidimensional vector space could have the similar meaning. The multidimensional vector space is created using word embedding technique [2, 9]. Then, semantically related words are organized into semantic vocabulary using word clustering technique. Words are clustered based on the similarity of words. Various information retrieval systems employed semantic vocabulary for query expansion [1–3]. In this work, we present Amharic information retrieval system with query expansion. Query expansion relies on the semantic vocabulary constructed automatically from a text corpus using word embedding technique followed by clustering of word senses.

The remaining part of this paper is organized as follows. Section 2 presents linguistic characteristics of Amharic. In Sect. 3, we present the proposed Amharic information retrieval system. Experimental results are discussed in Sect. 4 our conclusion is presented in Sect. 5.

2 The Amharic Language

2.1 Amharic Writing System

Amharic is the working language of Ethiopia. Although many languages are spoken in Ethiopia, it is serving as the *lingua franca* of the country. Amharic is written using Ethiopic script which has 34 characters with 7 vowels. The seven vowels are አ /ä/, ኡ /u/, ኢ /i/, ኣ /a/, ኤ /e/, እ /ə/ and ኦ /o/. Each character is modified with the vowels yielding seven orders. For example, the character ከ /kä/ is modified using vowels as ኩ /ku/, ኪ /ki/, ካ /ka/, ኬ /ke/, ክ /kə/ and ኮ /ko/. In addition, there are also labialized characters like ኳ /kua/, ሷ /sua/, ሏ /lua/, ሟ /mua/, etc. Ethiopic script uses punctuation marks such as comma (፣), semicolon (፤), full stop (።), colon (፥) and preface colon (፦). The script has its own characters for numbers as well.

2.2 Characteristics of Amharic Language

Amharic exhibits complex morphological processes through derivation and inflection that apply mainly on word classes such as verbs, nouns and adjectives [10, 11]. Typically, Amharic verbs are generated through a two-step process from verbal roots: *stem formation* and *verb formation*. Amharic verbal stems, from which various forms of verbs are formed, can be derived from verbal roots by affixing vowels. For example, the verbal stem ሰበር- /säbär-/ is derived from the verbal root

ስ-ብ-ር /*s-b-r*/. The process of Amharic verb formation is usually completed by marking stems for any combination of person, gender, number, case, tense/aspect and mood. Accordingly, the following verbs can be generated from the verbal stem: ሰበርኩ /*säbärku* 'I broke'/, ሰበርኩህ /*säbärkuh* 'I broke you'/, ሰበርን /*säbärn* 'we broke'/, ተሰበርኩ /*täsäbärku* 'I was broken'/, ሰበረች /*säbäräc* 'she broke'/, etc. The characteristic feature that a single instance of a verb can be marked for a combination of person, case, gender, number, tense, aspect, mood and others leads to the possibility of generating tens of thousands of verbs from a single verbal root through the processes of derivation and infection.

From the perspective of morphological structure, Amharic nouns can be derived and non-derived [10, 11]. Derived nouns are formed though morphological processes applied on various word origins. Words like ቤት /*bet* 'house'/, ሃገር /*hagär* 'country'/ and ሰው /*säw* 'human'/ are non-derived nouns. On the other hand, words like መልስ /*mäls* 'response'/ and ደግነት /*dägənät* 'generosity'/ are nouns derived from the verbal root ም-ል-ስ /*m-l-s* 'to respond'/ and the adjective ደግ /*däg* 'generous'/, respectively. Similar to nouns, Amharic adjectives can be derived and non-derived [10, 11] where derived adjectives can be formed from verbal roots by infixing vowels between consonants (e.g. ድ-ር-ቅ /*d-r-q* 'to dry'/ → ደረቅ /*däräq* 'dry'/), nouns by suffixing bound morphemes (e.g. ተራራ /*tärara* 'mountain'/ → ተራራማ /*täräramа* 'mountainous') and stems by prefixing or suffixing bound morphemes (e.g. ደካም- /*däkam-*/ → ደካማ /*däkama* 'weak'/). Amharic nouns and adjectives are inflected for number commonly by suffixing -አች /- *oc*/ or -ዎች /*-woc*/, definiteness by suffixing -ኡ /*-u*/ or -ዉ /*-wu*/, objective case by suffixing -ን /*-n*/, possessive case by suffixing different morphemes depending on the subject, and gender by suffixing -ኢት /*-it*/. These inflections can appear alone or in combination at the same time, along with prepositions and negation markers which leads to the generation of thousands of word forms from a single noun or adjective. For example, ያለባለቤቶቿ /*yaläbaläbetocu* 'without the owners of the house'/ is generated from the morphemes *yä-ʔalä-balä-bet-oc-u* (*yä-* 'preposition- of/with', *ʔalä-* 'nega- 'negation marker- not/without', *balä-* 'possessive 'possessive marker- owner of', *bet* 'house', *-oc* 'plural marker', and *–u* 'definite marker - the') where the core morpheme is the noun ቤት /*bet* 'house'/.

3 The Proposed Solution

The proposed Amharic information retrieval system performs query expansion using semantic vocabulary. The semantic vocabulary is automatically generated from a collection of Amharic documents by applying text processing. The processed text is used for neural word embedding from which clustering is made based on the similarity of word senses. Then, the semantic vocabulary is constructed from clusters of related words. On the other hand, the processed text is used for generating index terms representing Amharic documents. From the users side, the same text processing task is performed on the queries. Then, query expansion is made using the semantic vocabulary. Finally, the search of relevant documents is made by matching expanded query terms against index terms. Figure 1 shows system architecture of the proposed Amharic information retrieval system.

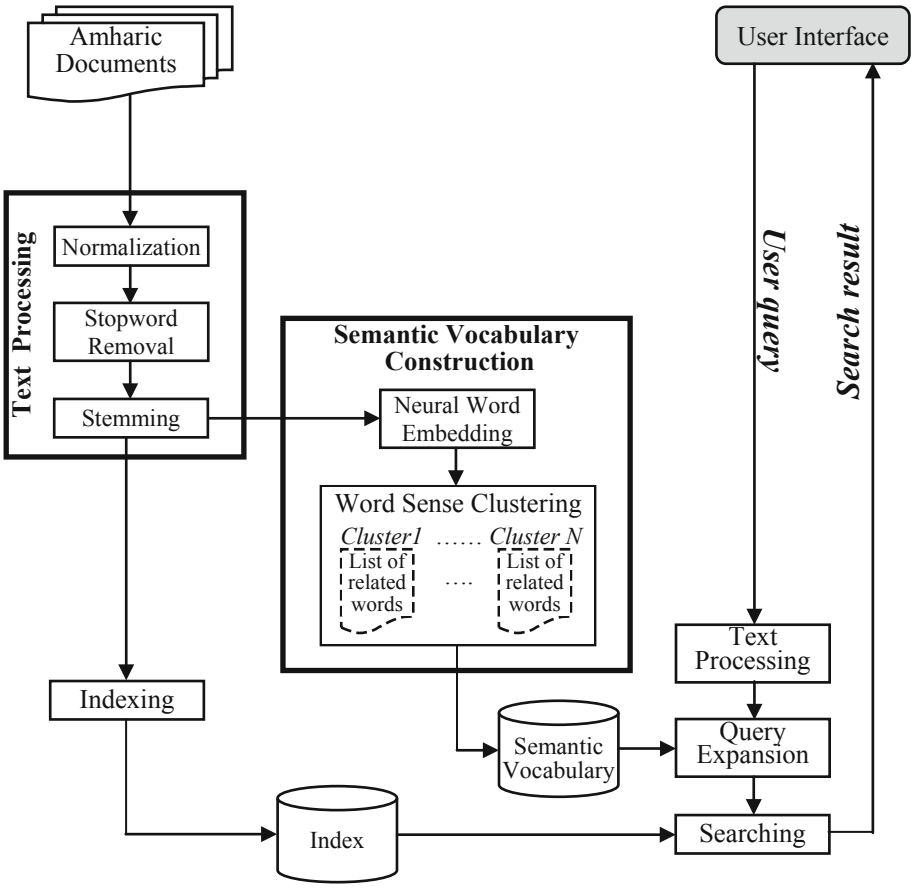

Fig. 1. System architecture

3.1 Text Processing

In general, information retrieval systems undergo text processing for a better document representation. However, the specific tasks may vary due to differences in the characteristics of languages. In our work, text processing involves character normalization, stopword removal and stemming. Text processing is performed on Amharic document collection during indexing and construction of semantic vocabulary. Similarly, it is also applied on user query.

Text Normalization. Text normalization is an essential step in Amharic text processing. Amharic is known to have a few set characters with similar pronunciation but different symbols. The base characters having such properties are: {ሀ /hä/, ሐ /hä/, ኀ/hä/, and ኸ /hä/}, {ሰ /sä/ and ሠ /sä/}, {አ /ʔä/ and ዐ /ʔä/}, and {ጸ /ṣä/ and ፀ /ṣä/}. Furthermore, the fourth orders of {ሀ /hä/, ሐ /hä/, ኀ /hä/, and ኸ /hä/} and {አ /ʔä/ and ዐ /ʔä/} have similar pronunciation with the respective base forms. As there are no standards established on how to use characters, a single word

may be written differently. For example, ፀሐይ /ṣähäy 'sun'/ may also be written as ጸሐይ /ṣähäy/, ፀሀይ /ṣähäy/, ፀኃይ /ṣähay/, ፀሓይ /ṣähay/, ጸሃይ /ṣähay/, etc. where all of them are considered as the same word (and pronounced the same). Thus, to avoid such variations in a word, we applied character normalization where ሀ /hä/, ሰ /sä/, አ /ʔä/ and ፀ /ṣä/ are used to represent the respective characters with similar pronunciation. We also use the first orders of characters to represent the respective fourth orders having similar pronunciation.

Stopword Removal. Stopword removal is a typical step in information retrieval. The objective is to remove non-content-bearing terms from documents and queries. Although standard stopwords are available for various languages, there are no standard and ready-made stopwords for the purpose of Amharic information retrieval. In our work, we identified stopwords from the document collection by considering frequently occurring words across each document. We used a total of 1200 terms are identified as stopwords. Examples include ነዉ, ነበር, ሆኖም, እና, ገለፁ, ዘግበዋል, አስታውቀ, ተናግረዋል, ብለዋል, ወደ, etc. A list of stopwords is created and stored in a file to filter out non-content-bearing terms during document and query processing.

Stemming. Stemming is required for conflating terms to a common form so as to avoid term variations arising as a result of morphological process. In our work, we modified the Amharic stemmer developed by Alemayehu and Willet [12] for stemming words.

3.2 Semantic Vocabulary Construction

The semantic vocabulary is used as a resource for query expansion. It is constructed automatically from the processed text. To this effect, we apply neural word embedding followed by clustering based on word senses. Neural word embedding vectorizes words on a multidimensional space using Word2vec based on the notion that words surrounding another word can be contextually similar. Word sense clustering automatically clusters vectors of similar terms across the *n*-dimensional spaces. The dimension of the word-space model is determined by the number of words to be plotted across the *n*-dimensional space. Dimension reduction is performed to get the most important features of the document.

Neural Word Embedding. The preprocessed and stemmed words are feed into the word-space modeling which embeds the words across a multidimensional space via the contexts of the words. The word embedding is implemented using Word2vec based on the continuous bag of words (CBOW) model. The CBOW model learns the embedding by predicting the current word based on its context. The idea behind CBOW architecture is that meaning can be inferred from a word surrounding another word. The Word2vec takes a text corpus as input and produces the word vectors as output. To vectorize the words, we consider the parameters like context window size, minimum occurrence count of words, dimension size, workers, architecture, mean value, hierarchical softmax, and negative sampling. To improve the overall performance of the system, we reduce the vector size to 300. The maximum distance between the target and the context word is considered to be 10. Training is carried out using parallel processing. Then, we apply

vector normalization for better performances and memory use. It implies computing the semantic distances between vectors by forgetting the previously learned vectors. Figure 2 shows sample word vector visualized using *t*-distributed stochastic neighbor embedding (*t*-SNE).

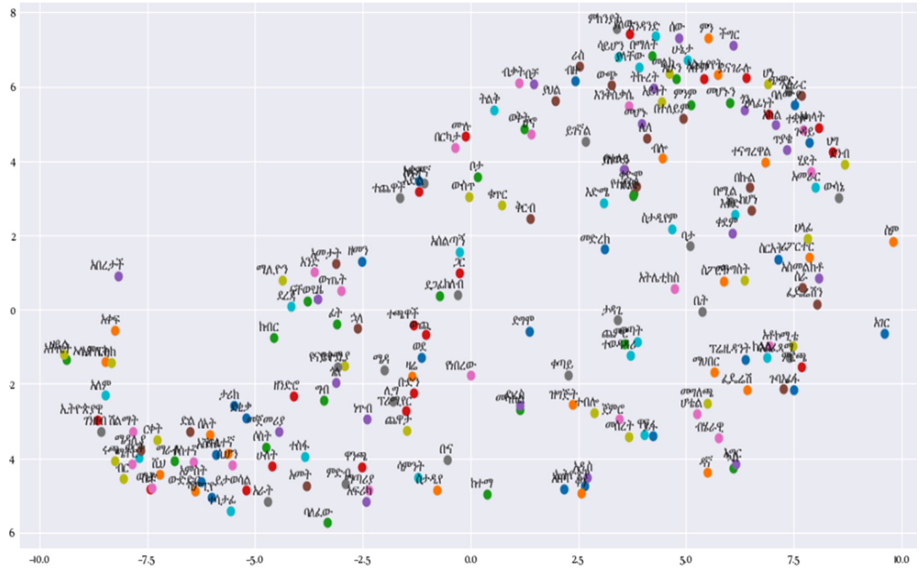

Fig. 2. Visualization of word vector

Word Sense Clustering. In order to construct the semantic vocabulary, semantically related words are clustered based on their similarity. The similarity of words is computed based on the cosine values which are implemented with vector orientation. Higher cosine values (approaching 1) indicate higher word relatedness. Based on their orientation words close to each other are grouped into a certain semantic cluster across the *n*-dimensional spaces without predefined labels or categories. An example of clusters of words in technology domain is shown in Fig. 3.

3.3 Searching

Searching of relevant documents is performed by matching index terms with expanded query terms. Index terms represent Amharic documents whereas expanded query terms represent users' needs. During the process of indexing, text processing activities of normalization, stopword removal and stemming are applied to Amharic documents. We also apply similar text processing on the queries. When the user provides the queries, the system expands the original query automatically from the semantic vocabulary. The queries are reformulated and words are matched against the indexed data which will return more relevant documents for the user based on the score for query terms and document ranks.

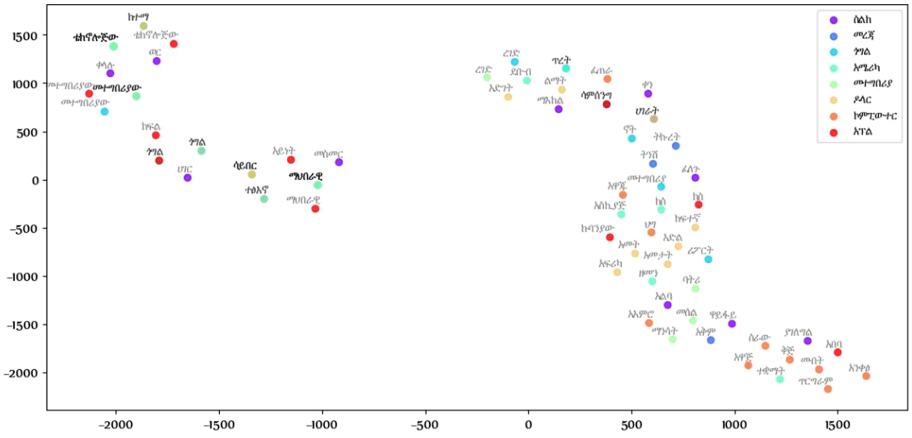

Fig. 3. Cluster of semantically related words in technology domain

4 Experiment

4.1 Corpus Collection and Implementation

Amharic text documents are collected from various sources covering 9 domains and another general domain. A total of 8,540 Amharic documents are used for training and testing purpose. Table 1 shows the domains long with the corresponding number of documents.The proposed information retrieval system is implemented using Python. We have constructed a total of 44,497 lists of semantically related vocabulary from all domains. Indexing and searching are implemented using Whoosh library.

Table 1. Amharic corpus used for training and testing.

Domain	Number of documents
Religion	1,189
Business	1,317
Sport	900
Politics	1,002
Law	1,037
Art	1,016
Health	272
Technology	978
Health	529
General	300
Total	**8,540**

4.2 Test Result and Discussion

For testing purpose, we use 9 queries and a total of 90 relevant and 518 indexed documents expanded with the top 5 number of similar words. When increasing the number of expanded words, the probability of returning the relevant documents increases. As a result, the recall increasing while precision is decreasing. This is expected due to the increasing number of retrieved documents. Relevance is measured using both ranked and unranked sets of retrieval. Table 2 presents mean average recall and precision for ranked and unranked retrieval sets evaluated with and without the use of semantic vocabulary for query expansion.

Table 2. Evaluation of the proposed system

Retrieval Set	Without semantic vocabulary		With semantic vocabulary	
	Recall (%)	Precision (%)	Recall (%)	Precision (%)
Unranked	69	44	84	24
Ranked	45	81	69	77

Generally, our system is better for recall-oriented applications as recall increases with the use of semantic vocabulary. The use of semantic vocabulary for query expansion helps the system perform better with the top five most similar terms. Figure 4 shows a comparison of retrieval effectiveness for unranked retrieval set with and without the use of semantic vocabulary for query expansion. The values in each test case show recall with semantic vocabulary (Rwsv), precision with semantic vocabulary (Pwsv), recall without semantic vocabulary (Rwosv), and precision without semantic vocabulary (Pwosv).

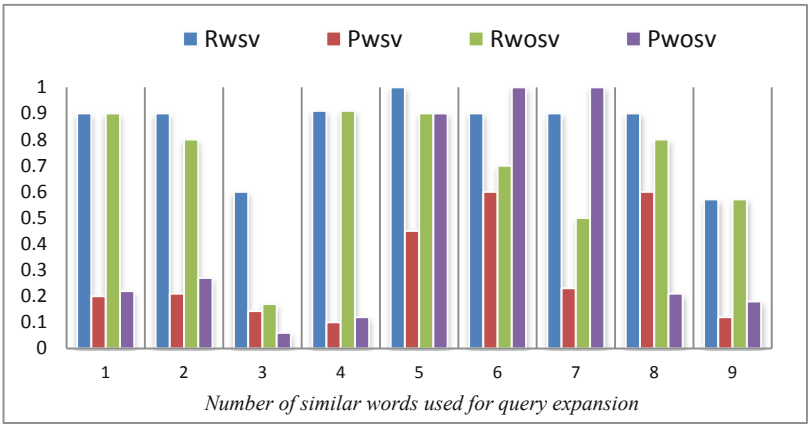

Fig. 4. Retrieval effectiveness of the proposed system

We show that ranked retrieval is generally better than unranked retrieval. Using ranked retrieval many relevant documents are retrieved in their priority relevance based

on the ranks in descending order. When the indexed documents are too huge ranked retrieval could save searching time than an unranked retrieval system. For ranked retrieval, we compare the performance of three ranking functions: *TF-IDF*, *BM25F* and *frequency*. Experimental results show that the BM25F ranking function outperforms better than other ranking functions.

5 Conclusion

The need to use and access lexical resources to enhance information retrieval using query expansion has become more fundamental in natural languages including Amharic. The resources made available for query expansion may be either manually built or automatically created from large document collections. Contextual meaning extraction also becomes more essential these days. Creating resources manually is labor-intensive, and time-consuming when the dataset is huge. Thus, we proposed an automatic way of creating Amharic semantic vocabulary using neural word embedding which helps to extract contextual meaning. The semantic vocabulary is used for query expansion where experimental results show significant improvement in recall albeit at the expense of some precision. Thus, future work is directed at employing additional techniques for improving precision. Moreover, we also recommend to deal further with the morphological characteristics of Amharic language.

References

1. Raza, M.A., Mokhtar, R., Ahmad, N., Pasha, M., Pasha, U.: A taxonomy and survey of semantic approaches for query expansion. IEEE Access **7**, 17823–17833 (2019)
2. Carpineto, C., Romano, G.: A survey of automatic query expansion in information retrieval. ACM Comput. Surv. **44**(1), 1–50 (2012)
3. Rivas, A.R., Iglesias, E.L., Borrajo, L.: Study of Query Expansion Techniques and Their Application in the Biomedical Information Retrieval, Sci. World J. (2014)
4. Aklouche, B., Bounhas, I., Slimani, Y.: Query expansion based on NLP and word embeddings. In: Proceedings of the Twenty-Seventh Text Retrieval Conference (TREC 2018), Gaithersburg, Maryland, USA (2018)
5. Kuzi, S., Shtok, A., Kurland, O.: Query Expansion Using Word Embeddings. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 1929–1932, (2016)
6. Venelin, K., Maria, S., Martí, M.: Comparing distributional semantics models for identifying groups of semantically related words. Procesamiento del Lenguaje Natural **57**, 109–116 (2016)
7. Alessandro, L.: Will Distributional Semantics Ever Become Semantic? In: Proceedings of the 7th International Global WordNet Conference, Tartu, (2014)
8. Claveau, V., Ewa, K.: Distributional Thesauri for Information Retrieval and vice versa, in Language and Resource Conference, LREC (2016)
9. Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.): ECIR 2019. LNCS, vol. 11437. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-15712-8
10. Assabie, Y.: Development of amharic morphological analyzer using a hybrid approach. Technical Report, Ethiopian Ministry of Communication and Information Technology, Addis Ababa, Ethiopia (2017)

11. Abate, M., Assabie, Y.: Development of amharic morphological analyzer using memory-based learning. In: Przepiórkowski, A., Ogrodniczuk, M. (eds.) NLP 2014. LNCS (LNAI), vol. 8686, pp. 1–13. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10888-9_1
12. Alemayehu, N., Willett, P.: Stemming of Amharic words for information retrieval. Literary Linguist. Comput. **17**(1), 1–17 (2002)