Amharic Text Clustering Using Encyclopedic Knowledge with Neural Word Embedding

Dessalew Yohannes
Department of Computer Science, Jimma University, Jimma, Ethiopia,
dessie.yohannes@ju.edu.et / abushay123@gmail.com /

Yaregal Assabie
Department of Computer Science, Addis Ababa University, Addis Ababa, Ethiopia,
yaregal.assabie@aau.edu.et

Abstract
In this digital era, almost in every discipline people are using automated systems that generate information represented in document format in different natural languages. As a result, there is a growing interest towards better solutions for finding, organizing and analyzing these documents. In this paper, we propose a system that clusters Amharic text documents using Encyclopedic Knowledge (EK) with neural word embedding. EK enables the representation of related concepts and neural word embedding allows us to handle the contexts of the relatedness. During the clustering process, all the text documents pass through preprocessing stages. Enriched text document features are extracted from each document by mapping with EK and word embedding model. TF-IDF weighted vector of enriched feature was generated. Finally, text documents are clustered using popular spherical K-means algorithm. The proposed system is tested with Amharic text corpus and Amharic Wikipedia data. Test results show that the use of EK with word embedding for document clustering improves the average accuracy over the use of only EK. Furthermore, changing the size of the class has a significant effect on accuracy.

1 Introduction
Text document clustering aims to find out common representative information from text documents and clustering these documents into the most relevant classes. Text clustering groups documents in an unsupervised way and there is no label or class information. Clustering approaches have to discover the connections between documents, and documents clustered based on these connections. Grouping documents into clusters is a basic step in many applications such as indexing, retrieval and mining of data on the web. Given huge volumes of documents, a good document clustering method may organize those huge numbers of documents into meaningful groups, which enables further browsing and navigation to be much easier. Traditionally, clustering of documents has been regarded as grouping them using predefined classes based on supervised learning techniques. The techniques used mainly employ features like words, phrases, and sequences from documents based on the frequency of features to perform categorization to predefined classes. However, such results are considered as unsatisfactory since huge volume of documents may not necessarily reflect predefined topics. Furthermore, recent trends show the need to shift to unsupervised learning where classes are to be constructed dynamically based on the semantics of their contents. In such cases, knowledge bases are used to augment unsupervised learning. Wikipedia is free online encyclopedia, which has become the largest electronic knowledge repository on the web with millions of articles contributed collaboratively by volunteers (Jens Lehmann et al., 2015). In Wikipedia, each article describes a single topic. Equivalent concepts are grouped together by redirected links and each article belongs to at least one category. Wikipedia makes much of its content available for offline analysis through dumps of its database (Wikipedia, 2020). These database dumps are commonly used as test-bed in the research community and numerous applications, algorithms and tools have been built around or applied to Wikipedia (Xiaohua Hu et al. 2009).

Word embedding is a modern approach for feature learning techniques in natural language documents. Word embedding is built on the idea that the semantics of a word arise simply from its
context (Tomas Mikolov et al. 2013). It captures both semantic and syntactic information of words, and can be used to measure word similarities, which are widely used in various natural language processing tasks. Accordingly, the features of word vectors representing contextual similarities between words can be manipulated arithmetically just like any other vector. In order to enhance Amharic text document clustering by leveraging semantics, two issues need to be addressed: a background encyclopedic knowledge base which can cover the relevant domain of individual document collections as completely as possible; and a suitable text feature extraction method which can enrich document representation by fully leveraging semantic terms, contexts and relations between terms. Thus, our hypothesis is that the inference abilities of encyclopedic knowledge coupled with the power of neural word embedding can be used to extract enriched text document features creating more accurate cluster of text documents. This study is an initial attempt to explore the use of encyclopedic knowledge with neural word embedding for clustering Amharic text documents. Moreover, unsupervised method of text document clustering is employed by using advantages of encyclopedic knowledge and word embedding for feature extraction that has not been included in the previous studies. The remaining part of this paper is organized as follows. In Section 2, the linguistic features of Amharic are discussed. Related works are presented in Section 3. Section 4 discusses the proposed system. Experiments are presented in Section 5, and conclusions are made in Section 6. References are provided at the end.

2 Amharic Language

Amharic Writing system: Amharic is the official working language of Ethiopia, currently having an estimated population of over 110 million (internet usage statistics, 2020). Due to its official status, Amharic is the lingua franca of Ethiopia and widely spoken throughout the country. It is also the second most widely spoken Semitic language, next to Arabic (Meron Sahlemariam et al., 2009). Many literary works, official documents and educational materials are produced using Amharic signifying its importance in the country. Amharic uses Ethiopic alphabet for writing where the alphabet is conveniently written in a tabular format of seven columns. The first column represents the basic form with vowel _reserve /i/ and the other six orders are derived from it by more or less regular modifications indicating vowels in the order of _reserve /u/, _reserve /a/, _reserve /e/, _reserve /o/, and _reserve /o/. For example, the base character _reserve /nā/ has the following modifications: _reserve /n/ā/, _reserve /n/ā/, _reserve /n/ā/,  Reserve /n/ā/, _reserve /n/ā/,  Reserve /n/ā/,  Reserve /n/ā/. Although the Ethiopic alphabet has many more characters, Amharic commonly uses 34 base characters along with modifications on the respective base characters. Furthermore, labialized characters like &/a/, .reserve /a/, .reserve /a/, .reserve /a/, etc. are derived from the base characters. The language also uses punctuation marks such as #: (full stop), , (comma), ; (semicolon), (colon), : (preface colon), etc.

Amharic Morphology: Like other Semitic languages, Amharic exhibits complex morphological processes through derivation and inflection that apply mainly on word classes such as verbs, nouns and adjectives [17, 18]. 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 _reserve /sābār-/ is derived from the verbal root _reserve /s-/āb-/reserve/. Moreover, verbal stems (e.g. _reserve /sābār-/ reserve/) can be derived from other verbal stems by affixing morphemes (e.g. _reserve /tāsābār-/ reserve/). 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: _reserve /sābārkə reserve I broke/, _reserve /sābārkə reserve I broke you/, _reserve /sābār reserve we broke/, _reserve /sābārkə reserve I was broken/, _reserve /sābārkə reserve she broke/, etc. This rich morphological feature exhibited on verbs leads to a case where marked verbs may represent complete sentences. For example, the word _reserve /alăm/tāsābār/ reserve 'he did not break us/', which is constructed from the morphemes al-sābār-ā-nə-m, is a complete sentence with the following linguistic information: al- reserve /-m/ /not/ , -sābār- reserve /did break/ , -ā- reserve /he/ and -nə- reserve /us/. 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 (Yaregal Assabie, 2017). Non-derived nouns mainly represent basic or primitive terms that refer to concepts, objects, entities, etc. whereas derived nouns are formed through morphological processes applied on various word origins. Words like ኤት /bet 'house/', ከር /hāgär 'country/' and ቢሮ /sāw 'human' are non-derived nouns. On the other hand, words like ወሮስ ሓ.RemoveAlls 'response' and እናት ስሮናት 'generosity' are nouns derived from the verbal root የወራት-ስት /m-l-s 'to respond/ and the adjective እናት /dāg 'generous/', respectively. In general, Amharic nouns can be derived from verbal roots, adjectives, nouns, stems and stem-like verbs by affixing vowels or bound morphemes. Similar to nouns, Amharic adjectives can be derived and non-derived (Yaregal Assabie, 2013) where derived adjectives they can be formed from verbal roots by infixing vowels between consonants (e.g. የትራት-የው /dārāq 'dry/'), nouns by suffixing bound morphemes (e.g. ፈዳት ᑈትራእ /tērām-a /tärarama 'mountainous') and stems by prefixing or suffixing bound morphemes (e.g. ከወታት-ማት /dākama- ከወታት /dākama ታው 'weak/). Compound words can also form derived nouns and adjectives. Although the morphological process of derivation of nouns and adjectives is complex by itself, even more complexity arises from their inflections. Amharic nouns and adjectives are inflected for number commonly by suffixing ከነት-ተር /-u or ከጉት-የው /-wul, definiteness by suffixing ከተቡ- ቡሮ or ከተቡ- /-u/, objective case by suffixing ከታት /-n/, possessive case by suffixing different morphemes depending on the subject, and gender by suffixing ከታት- ቢሰ. 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, የማስቀስ ሚስት-የወት /yalābālābetu 'without the owners of the house/ is generated from the morphemes yā-alā--balā-betoc- u (yā- 'preposition - of/with', alā- 'negation marker - not/without', balā- 'possessive marker - owner of', bet 'house', -oc 'plural marker', and -u 'definite marker - the') where the core morpheme is the noun ኤት /bet 'house/.

Amharic Grammar: Like other languages, Amharic has phrases, of which verb phrases and noun phrases are the most common forms. Amharic verb phrase is constructed from sequences of words having a verb as the head word which is normally located at the end of the phrase. Likewise, Amharic noun phrase is formed from sequences of words having a noun at the end of the phrase as head word. For example, የማስቀስ ሚስት-የወት ከስማት-ትምህርት /yāṭamhart betu tämariwoc 'students of the school/' is a noun phrase having ከስማት-ትምህርት /tämariwoc 'students/' as a head word located at the end of the phrase whereas የስማት ከአልስበሩትም /bārun ከስላስበርሬም 'did not break the door/' is a verb phrase having ከአልስበሩትም /alsābārutm 'did not break' marked for the singular definite object bārun and plural indefinite subject/ as a head word located at the end of the phrase. Simple Amharic sentences have subject-object-verb grammatical pattern. The following example shows a simple sentence constructed from subject, object and verb.

የትምህርት ሕጉት ከስላስበርሬም። የስማት ከስላስበርሬም。

Students of the school did not break the door.

Here, the subject is Yāṭamhart betu tämariwoc, the object is bārun and the verb is ከስላስበርሬም. Thus, Amharic simple sentences can be syntactically viewed as a noun phrase followed by a verb phrase. On the other hand, the structure of Amharic complex sentences is more complex than that of English. In the case of English, subordinating conjunctions are commonly used to split independent and dependent clauses. However, in the case of Amharic, the dependent clause along with the subordinating conjunction could be part of complex phrase that includes the verb as head word from the independent clause. The following example shows construction of a complex sentence where double underlines indicate dependent clauses and single underlines indicate dependent clauses.

መምህሩ የማስቀስ ሚስት-የወት ከስማት-ትምህርት /mämḥaru yāṭamhart betu tämariwoc bārun የስልስበርሬም ከመምህሩ።

The teacher knew that students of the school did not break the door.

In the sentence, Mämhəru ከመምህሩ is independent clause and yāṭamhart betu tämariwoc bārun የስልስበርሬም is dependent clause having የስልስበርሬም (affixed with the verb alsābārutm) as a subordinating
conjunction. Syntactically, Māmhəru is a noun and yātəmhərt betu tämariwoc bärun ?andalsābärut ?awqual is a complex verb phrase having the verb ?awqual as a head word. At the same time, ?awqual is a verb for the independent clause and thus marked for Māmhəru i.e., third person singular masculine.

With all these complex dependencies spanning the sentence, Amharic verbs and nouns are marked for several linguistic features inducing complex morphological structure. Thus, most high-level natural language processing applications developed for Amharic heavily rely on the proper use of stemmer or morphological analyzer is welcome. As always, the respective call for papers is the authoritative source.

3 Related Work

Various approaches have been proposed over the years to solve the problem of text document organization (clustering and classification) for documents written in different natural languages such as Amharic (Yohannes Afework, 2011), Chinese (Pu Han et al., 2013, Mingyu Yao et al. 2012), English (Anna Huang et al., 2009, Alahmadi et al., 2013), Arabic ( Hanane Froud et al., 2013), Ukrainian (Andrey Kutuzov et al. 2013), etc. (Huang et al.) Proposed a system for clustering documents with active learning using Wikipedia as a background knowledge base. This study explores the semantic knowledge in Wikipedia for grouping of documents and enabling automatic clustering of similar documents. Wikipedia concepts are utilized to create concept-based representation of text document and supervised approach is used for training with active learning. Hu et al. used Wikipedia concepts and categories for text document representation. Two approaches are proposed for mapping concepts to the documents. The first is dictionary-based approach that employs exact-match technique where it maps topical terms present in documents directly to Wikipedia concepts. The second approach employs relatedness match in which instead of mapping Wikipedia concepts to each document directly, it builds the connection between Wikipedia concepts and each document based on the contents of Wikipedia articles. (Yang et al.) Proposed an approach for mining hidden concepts based on short text clustering using Wikipedia as background knowledge base. This work takes into account increasingly available short texts in social networking platforms. Documents are enriched by searching related concepts where Wikipedia concepts are identified in documents. After texts are enriched with Wikipedia knowledge, clustering is performed using bisecting k-means algorithm based on topics. This work shows that the use of Wikipedia as a resource for enriching texts improves the performance of community mining. Review of related works indicates that two major approaches are used for text clustering or classification: keyword-based approach and semantic or concept-based approach. Keyword-based approach uses a long list of words as vector space to categorize a given document to a predefined class. This approach is often unsatisfactory for two reasons (Cheng-Lin et al. 2014): first, it keeps the dimension of the data very high; second, it ignores semantics or important relationships between terms like synonyms or antonyms. Concept based approach employing ontology has two limitations: (i) it classifies to predefined categories of text documents only which does not consider different kinds of documents; (ii) the ontology is developed usually for specific domain resulting in limited coverage. Semantic structure (the meaning associated with linguistic units like words) provides access to a large inventory of structured knowledge (the conceptual system). Furthermore, encyclopedic knowledge is grounded in human interaction with others and the world around us that is contributed by any volunteer. The limitation of using only encyclopedic knowledge for feature extraction is the contextual semantics of the document. The meaning of text depends on the aspects of context in which the texts are made. Recently, there is a growing interest to use encyclopedic knowledge for enhancing text mining tasks. Furthermore, the emergence of word embedding technology has significantly improved the performance of different text analysis tasks. In view of this, we identify significant term features to represent original content using encyclopedic knowledge. We also enrich features using word embedding and reduce data dimension without losing essential information in the text. Although there are many initiatives on text document clustering, to our best knowledge, there is no prior work on unsupervised clustering using encyclopedic knowledge with word embedding. Thus, our objective is to design
a system that is capable of improving the performance of Amharic document clustering.

4 Proposed System

In this paper, we propose a system that combines encyclopedic knowledge (EK) with the neural network-based word embedding the purpose of which is to take advantages of the good features both have in semantic based text document clustering. In order to perform text document clustering based on semantic knowledge from Wikipedia with neural word embedding, we design a system having six main components: Preprocessing, Neural Word Embedding, Structured Concept Construction, Text Feature Extraction, Text Feature Enrichment, Feature Weighting, and Document Clustering. In preprocessing component, documents are represented into usable and identifiable format or structure. This component is designed by considering common preprocessing activities which depend on the characteristics of languages. Neural Word Embedding computes the similarities between terms in a text. Structured Concept Construction generates structured knowledge representation of Wikipedia conceptual topic labels that can be used for text clustering. The encyclopedic knowledge from Wikipedia contains the structured representation of Wikipedia categorical concept vocabularies and tree like relationships between these categorical concepts. Feature extraction and enrichment components are used to represent a document in a form that inherently captures the semantics of the text. This helps to reduce the dimensionality of text documents. Feature Weighting helps to represent text documents as vectors from which Document Clustering will perform the classification task by measuring the similarity or relatedness between text documents. The architecture of the proposed system is shown in Figure 1.

5 Experimentation

Corpus preparation: In this work, we have collected two types of Amharic data for experiment: (i) Amharic Wikipedia database dump that is structured and used as encyclopedic knowledge base; and (ii) Amharic text document corpus that are collected from various sources. We have used Amharic Wikipedia dump that consists of category and categorical link. Amharic text document corpus is collected from Amharic bible, news agencies, broadcasting media, online newspapers and magazines. We use different sources to make data heterogeneous and style independent. Out of many text documents available, 3885 are randomly selected for testing. Based on the text contents, these documents are categorized by five domain experts. Table 2 shows the domain class of texts along with the respective number of collected text documents used for testing the system. The classes of text documents are assigned category names by experts based on the contents of documents. Feature Enrichment: The meaning of a text depends on the aspects of context in which the text is used. In practice, contextual terms are those that appear frequently in a small number of documents but rarely in the other documents and tend to be more relevant and specific for that particular group of documents, and therefore more useful for finding similar documents. Feature enrichment using context is used to handle the contextual relation of conceptual features. Using word embedding, the proposed solution enhances feature F with concepts and relationships from encyclopedic knowledge, which are related to terms in F. This process enriches document features contextually. By mapping tree like concept category relationship, related concepts of a text document are extracted and added to a feature of a document. On the other hand, the context of each related feature is extracted and added to feature vector using. The context of each concept feature is also extracted and selected based on the frequency of appearance. In this study, we applied word embedding technique word2vec to obtain the fixed-length feature vector. The vector representations of words learned by word2vec model has been shown to carry semantic meanings. Based on the learning resultant vector from a collection of texts, we define contexts of related categorical concepts in a text as terms obtained by the interconnection of the different conceptual terms employed in texts. The interconnections between concept features in the document are used to find the most probable contextual term of the text document.
Implementation and testing: we have trained Amharic text document corpus with more than 1.8 million terms which resulted trained vector model. The output has a dimension of 1xV that represent one-hot encoding of a term where V is the vocabulary size. The relational operations between terms like distance and analogy were used for feature enrichment process. To evaluate the clustering results, the comparison is done between documents clustered using unsupervised method with that of ground truth manually grouped by experts. Precision, recall and accuracy were computed. These measures estimate whether the prediction was correct with respect to the ground truth. We have also conducted experiment without applying feature enrichment processes (finding and mapping contexts of related concepts, contexts of concept features using word embedding) for further analysis and result comparisons. Accordingly, we have tested text document clustering by only using the encyclopedic knowledge for feature extraction. Table 1 shows the result of clustering text documents with and without the use of word embedding for feature enrichment.
6 Discussion

The relatedness between clusters can be interpreted as the more text documents grouped incorrectly to specific class denotes the two clusters are more conceptually related. This represents that the two clusters have more conceptual terms or their relationships that allows incorrect clustering of these types of text documents. When we visualize the distribution during clustering process, the text documents in Technology and Politics categories are distributed in all classes with varying amount during clustering. The distribution of these documents show that Politics and Technology classes have common or related conceptual terms that have a similar effect in short text document clustering. There are more conceptual terms used in Politics and Technology that can be used in all categories of documents. These and other document distribution differences during clustering come due to the conceptual relationship between categories. For example, more Technology text documents are grouped incorrectly to Art that shows as Technology documents have more conceptually interrelated words with Art. On the other hand, the result shows Religious documents represented have no documents clustered on Technology class that shows less relationship between the two classes. There were 5.05% wrongly clustered text documents. When we look into these documents, the conceptual terms in the documents are closer to incorrect cluster representatives. Therefore, these documents were grouped under incorrect clusters. There were also limitations in the text such as spelling errors that lead to wrong categorization of documents. In Preprocessing module there are tokenization, stop word-removal, and stemmer which contribute to difficulties in the process of document clustering. The documents are tokenized based on the space in between terms in the text but not all list of phrasal terms in Amharic can be tokenized using space. Furthermore, we have used an Amharic stemmer that had accuracy of 95%. These limitations had their own implications during feature extraction and weighting which subsequently affects the performance of clustering text documents.

The performance of the system with respect to the number of clusters was analyzed. The evaluation was made by including word embedding component in the system. It was observed that the accuracy of the system generally increases as the cluster size decreases. Figure 2 shows the graphical representation of the performance of the system with respect to variation in cluster size.

| Class Name | EK without Word Embedding | EK with Word Embedding |
|------------|---------------------------|------------------------|
|            | Precision | Recall | Accuracy (%) | Precision | Recall | Accuracy (%) |
| Religion   | 0.98      | 0.99   | 98.99        | 0.98      | 0.99   | 98.99        |
| Politics   | 0.87      | 0.99   | 87.88        | 0.91      | 0.99   | 91.91        |
| Technology | 0.87      | 1.00   | 87.00        | 0.91      | 1.00   | 91.00        |
| Business   | 0.80      | 1.00   | 80.00        | 0.90      | 1.00   | 90.00        |
| Health     | 0.95      | 1.00   | 95.00        | 0.91      | 1.00   | 91.00        |
| Art        | 0.69      | 1.00   | 69.00        | 0.93      | 1.00   | 93.00        |
| Sport      | 0.94      | 1.00   | 94.00        | 0.97      | 1.00   | 97.00        |
| Total      | 0.894     | 0.99   | 90.29        | 0.94      | 0.99   | 94.95        |

Figure 2: performance with respect to variation in cluster size
7 Conclusion

Owing to its complex morphology, Amharic poses unique difficulties in the development of various natural processing applications. In this work, we propose techniques of unsupervised Amharic text document clustering by enriching text features using encyclopedic knowledge with word embedding technology applied on stems. Word2vec is neural network-based word embedding model that can establish similarities between terms, which are in our case stems. We enrich text document features using the contexts of related categorical concepts and most probable contextual word of concept features based on trained word2vec model. Text features are extracted for each Amharic text document using encyclopedic knowledge with neural word embedding technology. The importance of the feature is evaluated using text weighting process, after which spherical k-means algorithm is applied for clustering text documents. With the use of word embedding technology on stems, experimental results demonstrate that our system can achieve improved performances on text clustering task. Furthermore, the performance of the system will be improved significantly, as the encyclopedic knowledge base gets richer.

References

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Soren Auer, and Christian Bizer, DBpedia A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia, 2015, Journal of Semantic Web.

Wikipedia free encyclopedia. 2020. Retrieved October 25, 2020 from https://en.wikipedia.org.

Xiaohua Hu, Xiaodan Zhang, Caimei Lu, E. K. Park and Xiaohua Zhou, Exploiting Wikipedia as External Knowledge for Document Clustering, Semantic Scholar, June 2009.

Tomas Mikolov, Ilya Sutskever, and Kai Chen, 2013. Distributed Representations of Words and Phrases and their Compositionality, Cornell University Library, New York.

Towards Data Science, Word Embeddings, retrieved January21,2013,from https://towardsdatascience.com.

Internet Usage World Statistics, “Internet and Population Statistics 2020”, retrieved from http://www.internetworldstats.com/, Last accessed on, November 22, 2017.

Meron Sahlemariam, Mulugeta Libsie, and Daniel Yacob, 2009. Concept-Based Automatic Amharic Document Categorization, In Proceeding of the 15th Americas Conference on Information Systems.

Yohannes Aferwork 2011. Automatic Amharic Document Categorization: The Case of Ethiopian News Agency, Unpublished Masters Thesis, Department of Computer Science, Ethiopia.

Pu Han, Dong-bo Wang, and Qing-guo Zhao 2013. Chinese Document Clustering Based on Weka, Proceedings of the 2011 International Conference on Machine Learning and Cybernetics, Guilin, July 2011, Addis Ababa University.

Mingyu Yao, Dechang Pi, and Xiangxiang Cong, Chinese Text Clustering Algorithm Based on Kmeans, in International Conference on Medical Physics and Biomedical Engineering, 2012.

Anna Huang, David Milne, Eibe Frank, and Ian H.Witten, Clustering Documents with Active Learning using Wikipedia, Eighth IEEE International Conference on Data Mining, February 10, 2009.

Alahmadi, Arash Joorabchi, and Abdulhussain E. Mahdi, A New Text Representation Scheme Combining Bag-of-Words and Bag-of-Concepts Approaches for Automatic Text Classification, IEEE GCC Conference and Exhibition, November 20, 2013.

Andrey Kutuzov, Mikhail Kopotev, Tatyana Sviridenko, and Lyubov Ivanova, Clustering Comparable Corpora of Russian and Ukrainian Academic Texts: Word Embeddings and Semantic Fingerprints, Semantic Scholar, Apr 2016.

Hanane Froud and Abdelmonaime Lachkar, Agglomerative Hierarchical Clustering Techniques for Arabic Documents, Springer, Switzerland, September 2013. Fawaz S. Al-Anzi and Dia AbuZeina, Big Data Categorization for Arabic Text Using Latent Semantic Indexing and Clustering, International Conference on Engineering Technologies and Big Data Analytics, Bangkok, January 2016.

W. Mersehazen, Addis Ababa: Birhanena Selam Printing Press, 1934. mining hidden concepts Awgichew and Mesfin Getachew, Manual Annotation of Amharic News Items with Part-of-Speech Tags and its Challenges, in European Language Resources Coordination (ELRC), 2006.

Yaregal Assabie. Development of Amharic Morphological Analyzer Using a Hybrid Approach. Technical Report, Ethiopian Ministry of
Communication and Information Technology, Addis Ababa, Ethiopia, 2017.

Alemayehu, Nega, and Willett Peter, Stemming of Amharic Words for Information Retrieval, Literary and Linguistic Computing, January 2002.

Cheng-Lin Yang, Nuttakorn Benjamasutin, and YunHeh Chen-Burger, Mining Hidden Concepts: Using Short Text Clustering and Wikipedia Knowledge, IEEE 28th International Conference on Advanced Information Networking and Applications Workshops, 26 June 2014.