AMHARIC TEXT CLUSTERING USING ENCYCLOPEDIC KNOWLEDGE WITH WORD EMBEDDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Digital technologies have made very easy and cheap to generate, store and publish different kinds of data. In this digital era, almost in every discipline people are using automated systems that generate information represented in text format in different natural languages. As a result, there is a growing interest towards better solutions for finding, organizing and analyzing these text documents. In this paper, we propose a system that clusters Amharic text documents using Encyclopedic Knowledge (EK) with neural word embedding. EK enabled the representation of related concepts and neural word embedding allowed as us to handle the contexts of the relatedness. During the clustering process, all the text documents pass through pre-processing stages. The enriched text document features were extracted f.0rom each document through mapping with EK and trained word embedding model. Finally, text documents are clustered using popular spherical K-means algorithm. In order to experiment the feasibility of the proposed system, Amharic text corpus and Amharic Wikipedia data were used for testing. The study shows that the use of EK with word embedding for Amharic text document clustering results improvement in average accuracy than that of using only encyclopedic knowledge. Furthermore, changing the size of the class has a significant effect on the rate of accuracy and shows that as the cluster size increases the gap in rate of clustering accuracy between using EK with and without word embedding increases

1 INTRODUCTION

The increasing amount of documents written in different languages, creates a need to manage that massive amount of varied information. Text document clustering is to find out the common representative information from the text documents and grouping these documents into the most relevant groups. Text clustering groups the document in an unsupervised way and there is no label or class information. Clustering approaches have to discover the connections between the document, and then based on these connections the documents are clustered. Grouping of documents into clusters is a basic step in many applications such as indexing, retrieval and mining of data on the web. Traditionally, clustering of documents has been regarded as grouping them using predefined classes on the basis of supervised learning techniques. The techniques used mainly uses features like words, phrases, and sequences from the documents based on counting and frequency of the features to perform categorization to the predefined classes. However, such results are considered as unsatisfactory since the huge volume of documents may not necessarily reflect the 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 argument 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 [1]. It is much more comprehensive and up to date. 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 [2]. These database dumps are commonly used as a test-bed in the research community and numerous applications, algorithms and tools have been built around or applied to Wikipedia

[3]. Furthermore, the meaning of text also depends on the aspects of context in which the texts are made.Word embedding is a modern approach for feature learning techniques in natural language documents. Word embedding build on the idea that semantics of a word arise simply from its context [4, 5]. 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. Neural networks are a modern and emerging computational approaches which are revolutionizing the current data analytic tasks. In word embedding technology, words or phrases from the vocabulary are mapped to vectors of numeric values in which similar words are expected to be close in the vector space [4]. The good feature of word vectors of contextual similarities between words they 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 the document representation by fully leveraging semantic terms, contexts and relations between the terms. Thus, 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 by using advantages of encyclopedic knowledge and word embedding for feature extraction that was not included in the previous studies was employed.

1.1 AMHARIC LANGUAGE

Amharic is an official working language of Ethiopia and the second most widely spoken Semitic language, next to Arabic [6, 7]. Amharic differs from structure of Semitic languages, especially in syntax. Amharic took the whole Geez alphabet and use it in the writing system. The Amharic alphabet does not have capital and lower case distinctions. It uses a unique script called fidel which is conveniently written in a tabular format of seven columns. Amharic has 34 base characters and total of 435 Ethiopic script characters. Like other Semitic languages, Amharic is one of the most morphologically complex languages [14]. Amharic nouns are the main carriers of information which can be grouped into derived and nonderived nouns.

2 PROPOSED SYSTEM

In this paper we propose combining encyclopedic knowledge (EK) with the neural network based word embedding 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 considered that the system would have six main components: Structured concept construction, text preprocessing, neural word embedding, text feature extraction, text feature enrichment, and feature weighting and clustering modules. The EK component contains the structured representation of Wikipedia categorical concept vocabularies and tree like relationships between these categorical concepts. In text preprocessing component, text documents are represented to usable and identifiable format or structure. This component is designed by considering common preprocessing activities and it will be modified depending on language structure. Feature extraction and enrichment component is used to represent a document in a form that it inherently captures semantics of the text. This would help to reduce dimensionality of the text document. Wikipedia (online encyclopedia) dump categorical concept relatedness is structured in the form of tree from Wikipedia categorical topic label successive links. Concept-category related feature is formed by mapping extracted conceptual feature of a text document with the structured categorical concept tree of the encyclopedic knowledge. Related feature is a list of related categorical concepts of the concept features. In this study, we defined the more related link of Wikipedia concept to be the word or phrase which is an immediate internal or external link of a particular conceptual topic label. 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.

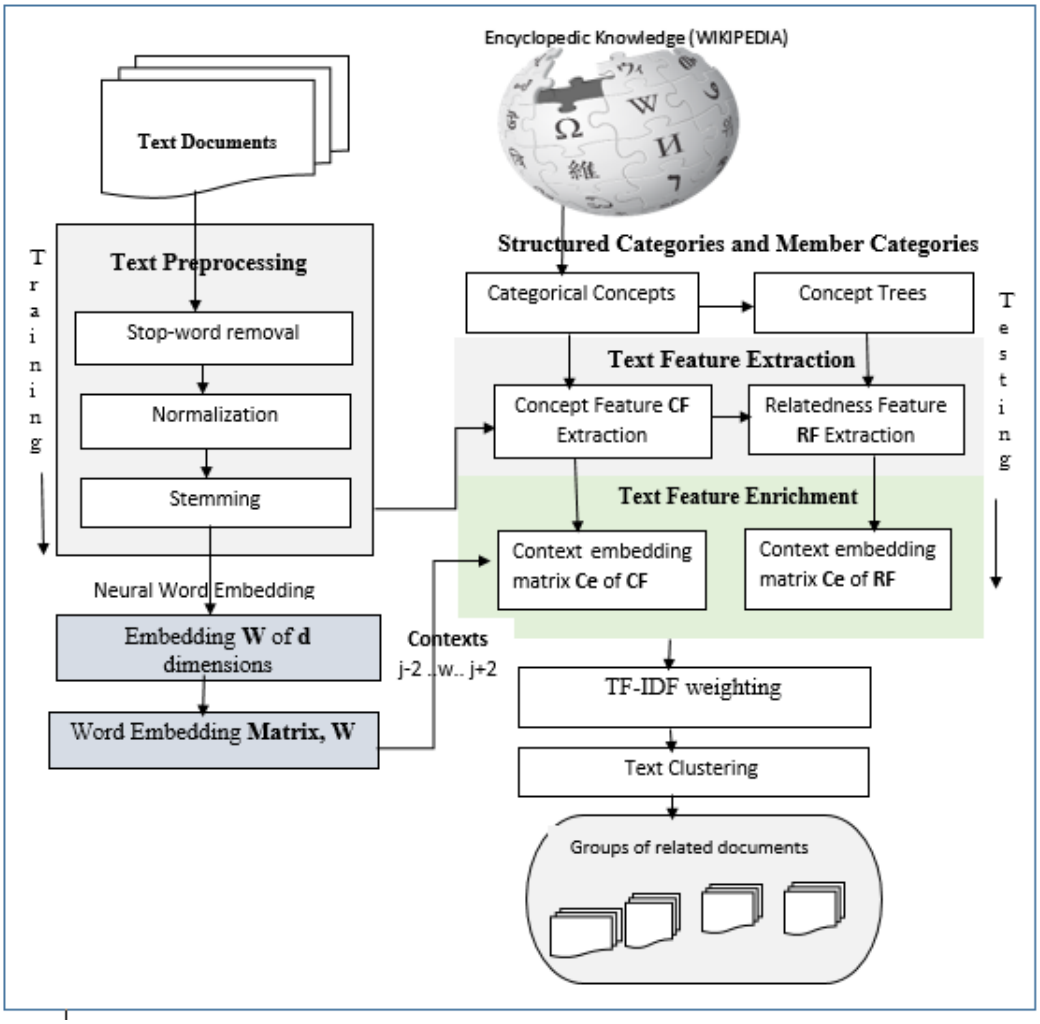


Figure 1: Architecture of Amharic Text Clustering using Encyclopedic Knowledge with Word Embedding

On the other hand, the context of each related feature is extracted and added using trained word embedding vector model. Text feature weighting and clustering component assigns numeric value for extracted features that is used to measure similarity or relatedness between text documents during clustering. The Architecture of the proposed system is shown in Figure 1.

3 Corpus Preparation

In this work, we have collected two types of Amharic data for experimentation. (1) Amharic Wikipedia database dump that is structured and used as encyclopedic knowledge base, and (2) Amharic text document corpus that are collected from different categories of documents for experimentation. We have used Amharic Wikipedia dump available on the date June 3, 2018 consists of category and categorical link. Amharic text document data are collected from Amharic bible, news agencies, broadcasting media, online newspapers and magazines. We use different sources to make data heterogeneous and writer independent. Out of many text documents available, 3885 are randomly selected for testing. Based on the text contents are discussing about, these documents are categorized manually by domain expert (3 journalists and 2 Amharic language teachers that are graduated in Amharic language). We have selected experts having a long time work experience. We have collected 3,885 text documents from 7 different categories. Thus groups of text documents

have manually assigned category name by the experts based on the documents are discussing about. We have trained two times 8,658 number of Amharic text document corpus with more than 1.8 million words and resulted trained vector model. The output has dimension 1xV, where V is the vocabulary size, that represent one-hot encoding of a word. In order to use insight, the relational operations between words like distance and analogy was used for feature enrichment process.

4 EXPERIMENTAL RESULTS AND EVALUATION

To evaluate the clustering results, the comparison was done between the document clustered using unsupervised method with that of manually grouped by experts. Precision and recall were calculated. These measures try to estimate whether the prediction was correct with respect to the underlying true categories. Thus, we have also done experimentation by skipping feature enrichment processes (finding and mapping contexts of related concepts, contexts of concept features using word embedding) for further analysis and result comparisons, i.e., we have tested text document clustering by only using the encyclopedic knowledge for feature extraction. The precision, recall and accuracy are used for evaluation.

The measures in figure 2 shows whether the prediction of each text documents class as being in the same cluster was correct with respect to the underlying true categories. It shows the result of clustering text documents using only encyclopedic knowledge and using encyclopedic knowledge with word embedding for feature enrichment. In figure 2, TC column denotes correctly clustered number of documents for each class; MC column denotes the number of documents missed in each class during clustering process; FC column denotes the number of falsely or incorrectly clustered documents in each class. The precision (P), Recall (R) and Accuracy (A) values are evaluated for each category of document.

Class Name	*No of Input Documents*	*With Word Embedding*						*Without Word Embedding*					
		TC	*MC*	*FC*	*P*	*R*	**A In %**	*TC*	*MC*	*FC*	*P*	*R*	**A In %**
Religious	1060	1036	2	22	0.98	0.99	98.99	1039	2	19	0.98	0.99	98.99
Politics	760	691	1	68	0.91	0.99	91.91	662	1	97	0.87	0.99	87.88
Technology	815	747	0	68	0.91	1.0	91.00	705	0	110	0.87	1.00	87.00
Business	370	333	0	37	0.90	1.0	90.00	295	0	75	0.80	1.00	80.00
Health	200	182	0	18	0.91	1.0	91.00	189	0	11	0.95	1.00	95.00
Art	230	214	0	16	0.93	1.0	93.00	158	0	72	0.69	1.00	69.00
Sport	450	435	0	15	0.97	1.0	97.00	422	0	28	0.94	1.00	94.00
Total	**3,885**	**3638**	**3**	**254**	**0.94**	**0.99**	**94.95**	**3470**	**3**	**412**	**0.894**	**0.99**	**90.29**

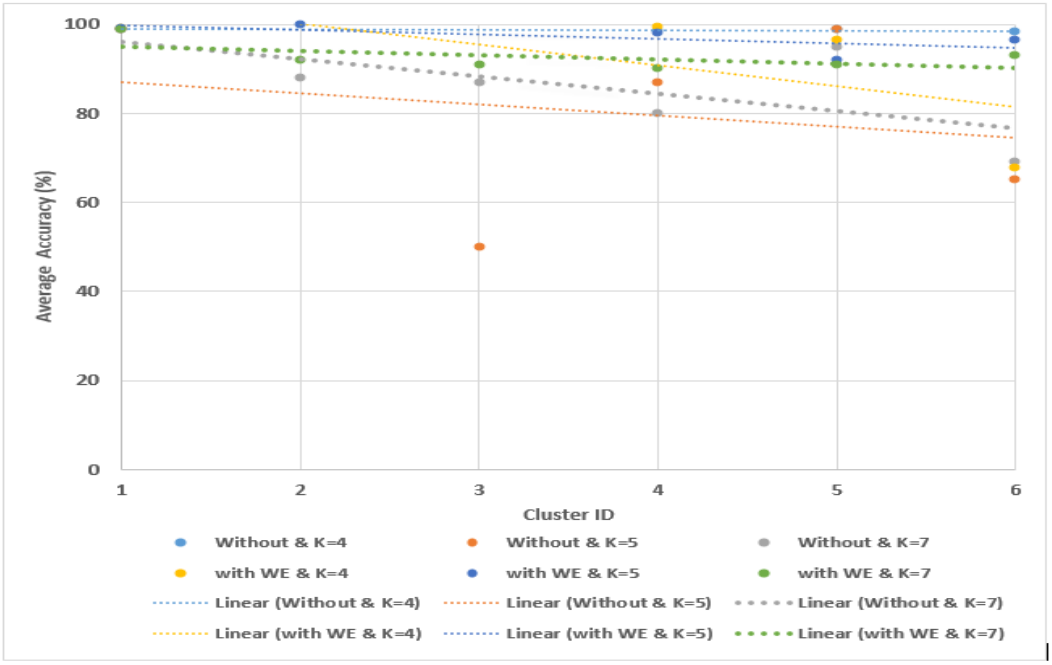
Figure 2: Evaluations of document clustering using EK with and without word embedding

When we visualize the distribution during clustering process, the text documents in technology category and politics are distributed in all classes during clustering in which the amount of distribution varies. The distribution of these documents show that politics and technology class has common or related conceptual terms that has the effect in short text document clustering. There are more conceptual terms used in politics and technology that can be probably used in all categories of documents. These and other document distribution difference during clustering comes due to the conceptual relationship between categories. For example, more technology text documents are grouped incorrectly to art that shows the technology class documents have more conceptually interrelated words with art that shows there is more relationship than other classes. On the other hand, the result shows

religious documents represented have no documents clustered on technology class that shows less relationship between the two class.

4.1 AVERAGE ACCURACY VALUES VS CLUSTER SIZE

For testing our proposed system, we used the cluster size K value 7 for seven types of text documents grouped by the experts. To see the effect of cluster size on the average accuracy value, we have tested by using 4, 5, and 7 kinds of documents from the collected corpus.



s l

Figure 3: The directions of Average Clustering Accuracy Vs Cluster Size

Figure 3, shows lines on a graph representing the general direction that a group of accuracy value points with different cluster size seem to be heading. In this table WE denotes with word embedding, without denotes without using word embedding and K represents the size of cluster. The cluster id 1, 2, 3, 4, 5 and 6 represents the cluster name religious, politics, technology, business, health and art respectively. The linear series shows the direction of Average accuracy of text document clustering by changing the size of cluster to 4, 5 and 7. The directions of Accuracy values are shown in both test cases, i.e., text documents clustering using encyclopedic knowledge with word embedding (WE) technology and clustering text documents using only encyclopedic knowledge for feature extraction. The linear trendline (linear series) shows the increasing or decreasing rate of accuracy that is a best-fit line to compare the two clustering results with change in the cluster size (K).

The difference of the two best-fit lines for cluster size 4, 5, and 7 are shown clearly in Figure 10. The gap between lines using K=7 (linear (with WE and K=7), linear (without and K=7)) shows the performance improvement of using EK with word embedding in which the rate of accuracy decrease significantly with increase in cluster size. Here we notice that changing the size of the class has a significant effect on the rate of accuracy.

The experimental results demonstrate that our system can achieve new improved performances on text clustering task. As shown on testing, these techniques and processes used in unsupervised text document clustering using encyclopedic knowledge with neural word embedding, the result of correctly clustered documents is 94.95 of the total test documents, which shows the average accuracy of the clustering is good. Furthermore, having enriched encyclopedic knowledge, improved word embedding technology and capability of linguistic preprocessing tools decides the final result of

clustering text documents. Thus as the encyclopedic knowledge base gets richer, the performance of the system will be improved significantly.

5 CONCLUSION

This research work had attempted to look into the techniques of unsupervised Amharic text document clustering by enriching text features using encyclopedic knowledge with word embedding technology. The context of a term in natural language text is given by the interconnection of the different words employed in a sentence for which the semantics of each word is known. Word2vec is neural network based word embedding model that can establish similarities between terms. In this study, we enriched 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. Uncorrelated features are eliminated to make more weighted concepts more descriptive by providing weight threshold. Among several k-means algorithms available, in this study, we used spherical k-means algorithm for clustering text documents. The experimental results demonstrate that our system can achieve new improved performances on text clustering task. As shown on testing, these techniques and processes used in unsupervised text document clustering using encyclopedic knowledge with neural word embedding, the result of correctly clustered documents is 94.95 of the total test documents, which shows the average accuracy of the clustering is good. Furthermore, having enriched encyclopedic knowledge, improved word embedding technology and capability of linguistic preprocessing tools decides the final result of clustering text documents. Thus the encyclopedic knowledge base gets richer, the performance of the system will be improved significantly.

6 REFERENCES

[1] 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, Journal of Semantic Web, 2015.

[2] Wikipedia free encyclopedia, retrieved from https://en.wikipedia.org, Last accessed on October 25, 2017.

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

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

[5] Towards Data Science, Word Embeddings, retrieved from https://towardsdatascience.com/, Last accessed on January 21,2018.

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

[7] Yohannes Afework, "Automatic Amharic Document Categorization: The Case of Ethiopian News Agency", Unpublished Master's Thesis, Department of Computer Science

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

[9] Mingyu Yao, Dechang Pi, and Xiangxiang Cong, "Chinese Text Clustering Algorithm Based on K-means", in International Conference on Medical Physics and Biomedical Engineering, 2012.

[10] 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.

[11] Alaa 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.

[12] 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.

[13] Hanane Froud and Abdelmonaime Lachkar, “Agglomerative Hierarchical Clustering Techniques for Arabic Documents”, Springer, Switzerland, September 2013.

[14] 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.