JOURNAL OF MULTIDIMENSIONAL RESEARCH & REVIEW

Website: http://www.jmrr.org Vol.1, Iss.1 Pages:31-

COMPARING K-MEANS AND EXPECTATION MAXIMIZATION ALGORITHM TO CLUSTER AMHARIC NOUN TERMS

Aklilu Mandefro Messele

Senior Researcher, Department of Computer Science,
Mega Computer and Research Center, Gondar, Ethiopia

Abstract

Natural Language processing, as a field of scientific inquiry, plays an important role in increasing computers capability to understand natural languages, the language by which most human knowledge is recorded. Many of Natural Language Processing (NLP) techniques have been used in information retrieval, though the result is not encouraging. This is because the value of information in the text usually is determined by nouns, to collect this information it should be detected first. This study is also believed to have significant contribution for researchers attempting to use noun terms for indexing on Amharic documents. It has been assumed by researchers that in text it is the noun terms that are content bearing elements. This study explores the application of unsupervised machine learning for constructing clusters of Amharic noun to this end comparative analysis is done between clustering using k-means and EM algorithm, that are used for identifying natural grouping based on a set of individual and a combination of high performing features .For consuming features forward feature selection approach is followed.

An experiment was conducted using 200 documents as the test set. This study identifies also noun term features such as Position, morphology, syntactic features and frequent information like (TF, IDF, and TF-IDF) that are capable of clustering nouns. According to the performance measure the accuracy by K-means is 80.75 %, whereas the accuracy by the Expectation Maximization /EM/ algorithm is 78.9 %. After the evaluation of the two algorithms, k-means clustering algorithm accuracy performance is better than EM Clustering algorithm. This research answers it is possible to use unsupervised learning for Amharic noun clustering. In doing so, this research finds that morphology, syntactic and positional information are features that are capable of distinguishing Amharic noun with other part of speech category (Verb, Adjective, Preposition and Adverb).

Keywords: *K-means, Expectation Maximization, Clustering, Amharic Noun, NLP.*

1. INTRODUCTION

Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text to do useful things. The goal of Natural Language Processing is to build computational models of natural language for its analysis and generation. These computational models provide a better insight into how humans communicate using natural language [1].

A lot of work has been done in the NLP community on clustering words according to their meaning in text [6]. The basic intuition is that words that are similar to each other tend to occur in similar contexts, thus linking the semantics of words with their lexical usage in text. One may ask why is clustering of words necessary in the first place? There may be several reasons for clustering, but generally it boils down to one basic reason[21]: if the words that occur rarely in a corpus are found to be distributional similar to more frequently occurring words, then one may be able to make better inferences on rare words. However, to unleash the real power of clustering one has to work with large amounts of text. The NLP community has started working on noun clustering on a few gigabytes of newspaper texts. But with the rapidly growing amount of raw text available on the web, one could improve clustering performance by carefully harnessing its power. Many of Natural Language Processing (NLP) techniques have been used in information retrieval, though the result is not encouraging. This is because the value of information in the text usually is determined by nouns, to collect this information it should be detected first.

The nouns in Amharic include notional words denoting subjects, objects, phenomenon, and also words used as objects of thought, any actions and states, features and relations. Because of this enormous diversity of reference, it is very useful to study nouns from the point of view of their formal characteristics. Whether a word belongs to the noun category can sometimes be determined according to morphological distinctions and sometimes syntactically (the place in the sentence or word-combination). In general, noun terms are defined as part of speech used for person's places objects and conditions [7]. To detect Noun terms from Amharic text unsupervised machine learning can be used for clustering noun terms.

Clustering is the process of grouping a set of objects into sets of similar objects. A cluster is a collection of data objects which are similar to one another within the same cluster and dissimilar to the objects in other clusters [2]. Clustering analysis helps to discover overall distribution patterns and relationship among data attributes. Clustering is applied widely in many areas, including pattern recognition, data analysis, image processing, and market research. Clustering can also be used in presenting search results [3].

There exist many clustering algorithms, which can be classified into several categories, including partitioning methods, hierarchical methods and density based methods [21]. A partitioning method classifies objects into several one level clusters. Each partition should contain at least one object. If each object belongs to only one clusters, it is called hard clustering; otherwise, it is called soft clustering. On the other hand, hierarchical methods create diagram that shows taxonomic relationship or decomposition of objects [2].Two approaches for building hierarchy are bottom-up and top-down. The bottom-up-approach, also called agglomerative approach, starting with n clusters containing one object; iteratively then merges objects into groups until finally only one group is left. The top down or divisive approach splits whole data set into several groups; there by iteratively split up clusters until every object is in only one cluster. A density based method introduced the notion of density, the number of objects in the "neighborhood" of an object. A given cluster continues growing until its density exceeds a threshold. Density-based methods can build clusters of arbitrary shape [4].

2. EXPERIMENTAL PROCEDURES, MATERIALS AND METHODS

2.1 The Amharic Writing System

The present writing system of Amharic is taken from Ge'ez. Ge'ez in turn took its script from the ancient Arabian language mainly attested in inscriptions in the Sabean dialect [55]. The original Sabaean alphabet is said to have had 29 symbols. When Ge'ez became the spoken and written language in common use in northern Ethiopia, it took only 24 of the 29 Sabaean symbols, modify most of them and add two new symbols to represent sounds of Greek and Latin loanwords not found in Ge'ez. The style of the writing was also modified to left to right. By the time Ge'ez ceased to be a living spoken and written language and replaced by Amharic and other languages, further changes

took place. Amharic did not discriminate in adopting the Ge'ez fidel, it took all of the symbols and added some new ones that represent sounds not found in Ge'ez[59]. These added alphabetic characters are ቸ, ጨ, ጀ, ኘ, ቨ, ሸ, ኸ, and ዠ.

Currently, the language's writing system contains 34 base characters each of which occurs in a basic form and six other forms known as orders. The seven orders represent syllable combinations consisting of a consonant following vowel.

This is why the Amharic writing system is often called syllabic rather than alphabetic, even if there is some opposition [59]. The 34 basic characters and their orders give 238 distinct symbols. In addition, there are forty others that contain a special feature usually representing labialization e.g.ቿ, ቋ. In Amharic there is no Capital-Lower case distinction. There are also punctuation marks and numeration system.

Amharic Numeration System

The Amharic numeration system consists of basic single characters for one to ten, for multiple of ten (twenty to ninety), hundreds and thousands. These numerals are derived from the Greek each has a horizontal stroke below and above. In the system, there is no symbol for representing zero value and it is not a place value system. In addition, the number system does not use commas or decimal points.

These situations make arithmetic computation using this system very complicated [7]. Both Amharic and Western numerals are in use today. Though the Amharic has long since been retired to a reserved use primarily for calendar dates and demarcation of sections in literature. Consequently, in most printed document Hindu- Arabic numerals are used.

Amharic Punctuation Marks

Amharic punctuation marks consist of as many as 10 punctuation marks in addition to the characters (Daniel, 1994). The basic punctuation marks are: the basic word divider, ሁለት ነጥብ-hulet netib, which has two dots arranged like a colon (፡) and a sentence ending is represented using, ዐራት ነጥብ- arat netib, four square dots arranged in a square pattern (።).Hulet Neteb is oddly used more in hand written practices today than in modern typesetting. Its place is almost completely taken over by space. Some others

equivalent to comma represented as, ነጠላ ሰረዝ--netela serez,(፣) and semicolon represented, ድርብ ሰረዝ-derib serez,(፤) and uses other Latin-based symbols like question mark(?), exclamation mark(!), quotes(“”) and parenresearch work().

2.2 Word Categories in Amharic

The linguistic characteristics of Amharic language have been studied by different researchers. The basic characteristic of the syntactic structure is SOV (Subject-Object-Verb). This structure SOV gives it a characteristic that modifiers in Amharic generally precede the words they modify. To start with, based on the way they categorize the Amharic words, research works in the area of Amharic word categorization can be grouped into two; namely, early and recent works [56]. In the early works such as [55]. Amharic words are categorized into the following eight categories (or parts of speech). These are the noun, Verb, Adjective, Adverb, Preposition, Pronoun, Conjunction and Interjection categories. In recent works such as [58], the early categorization of Amharic words is reduced into five, putting pronouns and conjunctions under the noun and preposition categories respectively. Here interjections, which are words without syntactic functions, are not considered as parts of speech at all.

Following the recent NLP works for Amharic(e.g. Atelach[11]; Mesfin[56]), this study adopts the classification by the early scholars but treating nouns and pronouns in the same part of speech class as suggested by Baye [58]. This preference is made because a direct implementation of a part of speech tagger developed by Mesfin[56], which is implemented as part of this study, adopts this kind of classification. Mesfin’s justification for such classification is that “the early categorization is more exhaustive and it allows the tagger to tag words exhaustively”, which found to be convincing.

2.3 System Architecture

Figure 2.1 present the system architecture of Amharic noun clustering. The process begin by applying data pre-processing (tokenization and Normalization) technique from given Amharic text collection, calculate term weighting using term frequency–inverse document frequency approach (TF*IDF), feature extraction then convert the data to CSV, which is suitable for the Weka package used for the automatic clustering,

and finally optimal features selected are used for evaluating the performance of clustering.

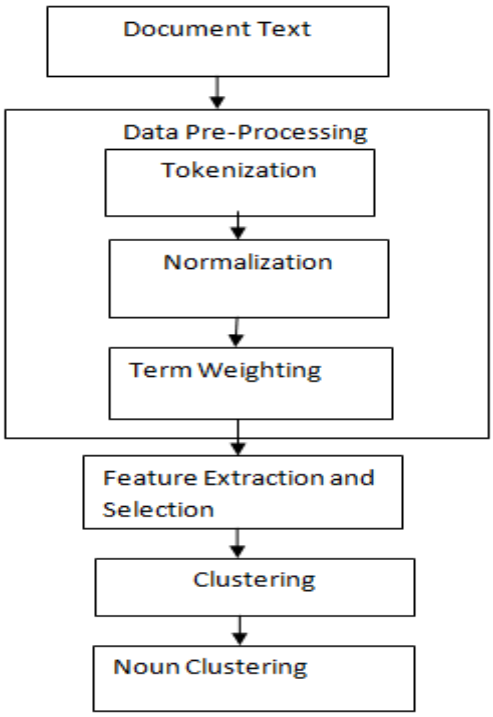

Fig 1 The system architecture in the general functions of the clustering prototype system [54]

2.4 Data Pre-processing

Amharic news articles are the main source of data set required for developing and testing clustering system. The fact that the news articles are easier to access, in sufficient amount and in electronic form, was the major motive to use them. Due to the requirement of larger time, finance and memory space in computation the number of test documents employed are 200 electronic Amharic news articles. The news articles are obtained from the Walta Information Center corpus.

Tokenization

Tokenization is the process of breaking a stream of text into words, phrases, symbols, or other meaningful elements called tokens. Texts in their raw form are just sequences of characters without explicit information about word and sentence boundaries. Before any further processing can be done, the training corpus is segmented into sentence and

then to word level. The sentence segmentation is done based on punctuation marks like (::?!). In Amharic text punctuations like (period:: question mark ? and exclamation mark !) are used to show sentence boundary.

Algorithm 1: Tokenization

1. initialize the variables to hold the word
2.read a character from the sentence (document)
3.check if the character is any one of the Amharic delimiter (punctuation mark, space, tab, carriage return, and line feed characters)
4.if not, concatenate the character to the variable
5.else if the character length is above one character report the word
6.If there is more data to process, go to step 1.

Normalization

The motivation for normalization is the observation that many different strings of characters often convey essentially identical meanings with different symbol. Given that to get at the meaning that underlies the words, it seems reasonable to normalize superficial variations by converting them to the same form. In this study, choosing one letter for the group of letters with the same sound and replace the remaining ones is taken as a solution to the problem. Therefore, if a character is any of ሓ, ኃ, ኻ, ሃ, ሐ or ኀ (all with the sound 'h') then, it is replaced by 'ሀ'. Also the different orders of ሐ and ኀ are changed to their corresponding equivalent orders of ሀ. Similarly, all orders of ሰ (with the sound 's') are changed to their corresponding equivalent orders of ሰ, all order of ዐ (with the sound 'a') are changed to their corresponding equivalent orders of አ, all orders of ፀ (with the sound 'tse') are changed to their corresponding equivalent orders of ጸ. The following algorithm is used for normalization.

Algorithm 2: Normalization

1. Read the character form the tokenized file
2. If the character is any of

ሃ/ኅ/ሐ/ኃ/ሓ then

Change it to ሀ Exit

Else if it is ሠ

Change it to ሰ Exit

Else if it is ዐ/ዓ/ኣ

Change it to ኣ

3. If the character that follows is a diacritic marking, attach it to the changed base character.

Term Weighting

Term weighting is used to select important terms. The techniques such as TF, IDF and TF*IDF are used to describe frequency of words [18]. In order to calculate the normalized frequency of term i in a document j first, count the number of occurrence of term i in a document j and divided by the total number of terms document j contains.

$$tfij = fij/\max(fik) \qquad (2.1)$$

Where tfij is a normalized term frequency of a given word, fij occurrence of term i in document j and maxfik stand for maximum number of terms in document.

After calculating normalized term frequency of each term, it is possible to find the inverse document frequency of a given word. In order to find IDF value first document frequency of term i must be defined. To find the document frequency of a given term, count the number of documents that contains the term i. Therefore, IDF of a given word calculated as follow.

$$IDF = tfij * \log(N/dfi) \qquad (2.2)$$

Where TF is the frequency of each term in the respective document and DF is the number of documents that contain the given term. So a combination of TF and IDF value of a word gives a way to include both frequent and infrequent words [13].

$$TF * IDF(w) = tfij * \log(N/dfi) \quad (2.3)$$

Where df_i is defines the number of documents in a collection that contains a term i, where N is the total number of documents in the collection.

2.5 Clustering Algorithms

Clustering is an unsupervised learning method that constitutes a cornerstone of an intelligent data analysis process. It is used for the exploration of inter-relationships among a collection of patterns, by organizing them into homogeneous clusters. Clustering has become one of the most active area of research and the development. Clustering attempts to discover the set of consequential groups where those within each group are more closely related to one another than the others assigned to different groups. The resultant clusters can provide a structure for organizing large bodies of text for efficient browsing and searching. There exists a wide variety of clustering

algorithms that has been intensively studied in the clustering problem. Among the algorithms that remain the most common and effectual, the iterative optimization clustering algorithms have been demonstrated reasonable performance for clustering, the Expectation Maximization (EM) algorithm, and the well-known k-means algorithm[16].EM and K-means are similar in the sense that they allow model refining of an iterative process to find the best congestion. In this paper, the performance of both algorithms, EM and K-means, for purity assessment of Amharic noun clustering is compared. Experimental results are analyzed and described by comparing two algorithms.

The k-means algorithm

The k-means algorithm is one of the best known partition clustering methods. The strategy it employs, detailed in Algorithm, consists essentially in iterating through the set of instances d1, . . . , dn assigning instances to the clusters with the nearest means (updating centroids), the cluster means and continuing the reassignments until a stopping (or convergence) criterion is met. A natural stopping criterion is to stop when no new reassignments take place. From various clustering algorithm that are based on minimizing an objective function k-means is one of the most studied and widely used clustering algorithm [16].Unlike hierarchical clustering, k-means starts off with a target number k of clusters and generates a flat set of clusters.

Algorithm 1: K-Means algorithm

K-means (x= $\{\vec{d}_{1,\ldots,}\vec{d}_{n}\}\subseteq R^{m}$, k) : 2^{R}

C: 2^{R} /*μ α *set of clusters* */

D:R^{m} x R^{m} $\rightarrow$ R /* *distance function* */

μ:$2^{R} \rightarrow$ R /* μ *computes the mean of a cluster*/

Select C with k initial center $\vec{f}_{1},\ldots\ldots\vec{f}_{k}$

While stopping criterion not true do

for all clusters cj $\in$ C do

cj $\longleftarrow$ $\vec{d}_{i} \mid \forall fid(\vec{d}_{i},f_{j}) \leq d(\vec{d}_{l},f_{t})\}$

done

for all means $\vec{f}_{j}$ do

$\vec{f}_{j} \leftarrow \mu(cj)$

done

done

return

The strategy it employs, detailed in Algorithm 2.1, consists essentially in iterating through the set of instances d1, . . . , dn assigning instances to the clusters with the nearest means (centroids), updating the cluster means and continuing the reassignments until a stopping (or convergence) criterion is met. A natural stopping criterion is to stop when no new reassignments take place. Unlike hierarchical clustering, k-means starts off with a target number k of clusters and generates a flat set of clusters.

The criterion used here is the distance between the instances. Though there are a number of options available as distance metrics, Euclidean distance is the most common choice[13]. The cluster mean is calculated as follows:

Euclidian distance measures a straight line distances between two points in Euclidean space [2].

If P and Q are two points in space then Euclidian Distance between these two points is measured as: summation of squared difference between them, as shown below in equation 2.4.

$$dis(p,q) = \sqrt{\sum_{i=1}^{n} \langle pi - qi \rangle^2} \qquad (2.4)$$

Other type of distance measurement is, Manhattan Distance which is the distance between two points measured along axes at right angles. This is also known as city block distance.

Manhattan Distance can be calculated as

$$D(p,q) = \sum_{i=1}^{n} |pi - qi| \qquad (2.5)$$

This study applied Euclidian distance to measures the distance between the instances. Because Euclidean distance is the most common choice for k-means algorithm and one advantage of this method is that the distance between any two objects is not affected by the addition of new objects to the analysis [13].

EM Algorithm

Expectation Maximization is a type of model based clustering method. It attempts to optimize the fit between the given data and some mathematical model. Such methods are often based on the assumption that the data are generated by a mixture of underlying probability distributions. It is iterative in nature and finds maximum likelihood solutions. With reference to [13], Expectation Maximization consists of two steps:

The maximization step finds the new clustering or parameters that maximize the expected likelihood in probabilistic model- based clustering.

Algorithm 4: EM Algorithm

Input: c: the number of clusters

D: a dataset containing n objects

Output: A set of k clusters

Method:

1) First find initial centers/centroids which will be the initial input.

2) Compute distance between each data point and each centroid using cosine distance formula or any other distance formula.

3) Assign weights for each combination of data point and cluster based on the probability of membership of a data point to a particular cluster.

4) Repeat

i) (Re) assign each data point to the cluster with which it has highest weight i.e., highest probability.

ii) If a data point belongs to more than one cluster with the same probability, then (re)assign the data point to the cluster based on minimum distance.

iii) Update the cluster means for every iteration until clustering converges.

2.6 Feature Extraction

For representing noun word class for clustering, extract syntactic, distribution and morphological information of Amharic words are extracted.

Syntactic Information

Amharic words usually placed in Subject-Object-Verb order. Therefore, there is considerable information to be gained from positions of words in a sentence. For example, noun in Amharic is always at the beginning and middle of the sentence. By considering the structural information this study filters out Amharic noun term.

Word Position

Positional information enables to predict the category of the word. In identifying word positional information in a sentence punctuation marks have a great role. In Amharic ending of a sentence most of the time is shown by a period (::), question mark (?) and exclamation mark (!) as a result any word which is appeared before those punctuation marks was counted as the last element of the sentence [60]. A word which is occur next to sentence bounder marks also consider as the first element of the sentence. In order to identify the mid position of a word the following mechanism is used, any word which is occurred between any two words is taken into account as a middle position of a word and this was done for every word.

Morphological Information

Words of different Part-of-Speech category have different affixes associated with them. In Amharic it is possible to differentiate words based on the prefix and suffix attached to them. For example, Amharic Noun plural most of the time attached with the suffix/-ኦች ooč/and /-ዎች wooč/. This morphological information serves as a filtering mechanism for mis-clustered word category by looking prefix and suffix attached with the word.

2.7 Data Conversion

After data preprocessing is complete and number of feature of noun term is known the next step is converting the dataset to a format appropriate for automatic clustering using Weka, which expects the source data for clustering to be in CSV format. This step

therefore involves the conversion of the pre-processed data to CSV, which is suitable for the Weka package used for the automatic clustering.

Weka processes data sets that are ARFF format.

ARFF format files

The following is an example of an ARFF file for a dataset:

@relation 'Morphology & Position & TF'

@attribute Morphology numeric

@attribute position numeric

@attribute tf numeric

@data

1,1,0.021053

It consists of three parts. The @relation line gives the dataset a name for use within Weka. The @attribute lines declare the attributes of the examples in the data. Each line specifies an attribute's name and the values it may take. In this example the attributes have numeric values so these are listed explicitly. The remainder of the file lists the actual examples, in comma separated format; the attribute values appear in the order in which they are declared above.

2.8 Feature Selection

Feature selection for clustering is the task of selecting important features for the underlying clusters. Commonly used heuristic methods for feature selection are forward or backward selection or some combination of both. A forward selection method first finds the best feature among all features, and then using the already selected features finds the next best two-component feature subset. Afterwards it moves to find the best triple out of all the combination of any three input features, etc. The subset that outputs the maximum purity is output as the best subset. Backward selection algorithm is the opposite of the forward selection algorithm. There are many other search techniques

that can be applied to feature selection. In the experiments we use a forward selection algorithm.

2.9. Evaluation technique

There are three approaches to study cluster validity [10]. The first is based on external criteria. This implies that evaluate the results of a clustering algorithm based on a pre-specified structure, which is imposed on a dataset, i.e. external information that is not contained in the dataset. The second approach is based on internal criteria. It may evaluate the results of a clustering algorithm using information that involves the vectors of the datasets themselves. Internal criteria can roughly be subdivided into two groups: the one that assesses the fit between the data and the expected structure and others that focus on the stability of the solution [22].The third approach of clustering validity is based on relative criteria, which consists of evaluating the results (clustering structure) by comparing them with other clustering schemes.

This study focuses on external clustering evaluation, i.e., evaluation against manually annotated gold standards, which exist for almost all such NLP tasks. Because external evaluation is the dominant form of clustering evaluation in NLP [30].The external validation measures are extremely useful in deducing the domain to which the clustering structure is ascertained by a clustering algorithm that matches some external structure. This is compared to the individual designated class labels. External validation measures criteria evaluate the final clustering output result with respect to a pre designated structure.

There are many external validation measures like Entropy, Purity, NMI Measure, F-Measure and other.

- F-measure

F-measure combines the precision and recall concepts from information retrieval.

$$F(i,j) = \frac{2Recall(i,j)Precision(i,j)}{Precision(i,j) + Recall(i,j)} \quad (2.6)$$

Where recall and precision for each class are calculated as:

$$Recall(i,j) = \frac{n_{ij}}{n_i} \text{ and } \quad Precision(i,j) = \frac{n_{ij}}{n_j} \qquad (2.7)$$

Were nij is the number of objects of class i that are in cluster j, nj is the number of objects in cluster j, and ni, is the number of objects in class j and class i is given by the following equation:

The F-Measure values are within the interval [0,1] and larger values indicate higher clustering quality. Normalized Mutual Information (NMI) is another measure of cluster quality, the NMI of two labeled objects can be measured as [25].

$$NMI(X,Y) = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}} \qquad (2.5)$$

Where, l(X, Y), denotes the mutual information between two random variables X and Y and H(X) and H(Y) denote the entropy of X,[X consensus clustering while Y the true labels].

Entropy is a measure to compute the entropy of a dataset, it's needed to calculate the class distribution of the objects in each cluster as follows [25]

$$E_j = \sum_i p_{ij} \log (p_{ij}) \qquad (2.6)$$

Entropy measures the purity of the clusters class labels. Thus, if all clusters consist of objects with only a single class label, the entropy is 0.However, as the class labels of objects in a cluster become more varied, the entropy increases.

Where the sum is taken over all classes. The total entropy for a set of clusters is calculated as the weighted sum of the entropies of all clusters, as shown in the next equation.

$$E = \sum_{i=1}^{m} \frac{\text{nj}}{\text{n}} \text{Ej} \qquad (2.7)$$

Were *nj* is the size of cluster *j*, *m* is the number of clusters, and *n* is the total number of data points.

Purity

To compute cluster purity, each group of lexicon was assigned to the class which is most common in the cluster, and then accuracy of this assignment is measured by counting the number of correctly clustered words divided by the size of cluster N [31].

We compute the purity for each cluster. For each cluster, the Purity $pj = \frac{1}{ni} Max_i (n_j^i)$ is the number of objects in *j* with class label *i*. In other words, p_j is a fraction of the overall cluster size that the largest class of objects assigned to that cluster represents. The overall purity of the clustering solution is obtained as a weighted sum of the individual cluster purities and given as:

$$Purity = \sum_{j=1}^{m} \frac{n_j}{n} P_j \quad (2.8)$$

where nj is the size of cluster j ,m is the number of clusters , and n is the total number of objects.

The entropy and purity are widely used measures. But purity is one of very primary validation measure to determine the cluster quality and it is a simple and transparent evaluation measure.

2.10. Experimentation

In conducting an experiment the experimental setting is classified into six based on the feature used to cluster. Each data set contains the same feature terms across the experimental setting but with different type of feature. The first dataset contains words with TF value. The second dataset is with, words and the corresponding IDF value. The third one includes TF*IDF value. The fourth one contains positional information in relation to words. The next dataset comes with words and syntactic features. The last experiment is done with the sequence of morpheme which is attached to words. This information is used as a filtering mechanism which assists to guess category of the word.

The clustering was done by using both K-means algorithm and the expectation maximization (EM) algorithm based only considering individual feature. Based on the evaluation of clustering purity result an optimal purity combination of those features is selected by applying forward feature selection method, which is taking a mixture of feature which registers high purity value this combination continues until purity drops, because computationally more efficient than backward elimination to generate nested subsets of variables and also it’s a very attractive approach, because it's both tractable

and it gives a good sequence of models [13]. Finally, optimal features selected are used for evaluating the performance of the clustering. In this study optimal feature refers to features that have maximum purity clustering value.

3. Results and Discussion

The Experiment was done by using both *K*-means algorithm and the expectation maximization (EM) algorithm based only considering individual feature. The best way of evaluating a feature selection method for clustering is to check the correctness of the selected features, i.e., how well the selected features match with the actual important features. According to this evaluation criterion, we first evaluate individual feature datasets. We used forward selection method and selected the optimal subset that has the overall maximum purity. Cluster evaluation methods that consider external criterion compare the result of the clustering algorithm against with some external benchmark is used to calculate the performance of the system. The standard which is used to evaluate the cluster quality is the Walta Information Center news which was annotated manually by the staff of Ethiopian Language Research Center (ELRC).

The experiment has six features and 4072 instance in 2 clusters.

Experiment	***Feature used for Representation***	***Number of cluster selected***	***Cluster purity (%)using k-means algorithm***	***Cluster purity (%)using EM algorithm***
First	TF	2	34.3	32.8
	IDF		34.8	33.2
	TF*IDF		35.2	34.5
	Syntactic feature		41.2	40.2
	Position		52	51
	Morphological		72	70

Table 1: Experiment, Individual feature representation, cluster purity using both k-means and EM algorithm.

To evaluate each feature subset forward selection method compare individual feature purity value and finds the best feature among all features. The subset that outputs the maximum purity is output as the best feature.

So the Result shows that among individual feature representation *Morphological feature* show maximum purity both in *k-means 72%* ***and*** *EM algorithm 70%***.** And then

using the already selected features finds the next best feature. The next experiment used the already selected features, which is morphology feature, and combines with each individual feature.

Experiment	***Feature used for Representation***	***Number of cluster selected***	***Cluster purity (%)using k-means algorithm***	***Cluster purity (%)using EM algorithm***
Second	Morphological –TF	2	72.8	70.8
	Morphological –IDF		73.2	71.6
	Morphological – TF*IDF		73.6	71.8
	Morphological – syntactic		76	74.8
	Morphological– Position		78	75.6

Experiment	***Feature used for Representation***	***Number of cluster selected***	***Cluster purity (%)using k-means algorithm***	***Cluster purity (%)using EM algorithm***
Third	Morphological– position-TF	2	78	77
	Morphological– position– IDF		78	77.5
	Morphological– position– TF*IDF		78	77.8
	Morphological– position– syntactic		80.75	78.9

Table 2: Experiment, Combined two feature representation, cluster purity using both k-means and EM algorithm

The second experiment result shows that *Morphological –Position feature* combination has maximum purity compared with other feature combination, which is k-means value and EM value. The next experiment used the already selected features, which is *Morphological –Position* feature, and combines with each individual feature.

Table 3: Experiment, combined three feature representation, cluster purity using both k-means and EM algorithm

The Third experiment result shows that *Morphological–position–syntactic feature* combination has maximum purity compared with other feature combination, which is k-means value and EM value. The next experiment used the already selected features, which is *Morphological –Position–syntactic* feature, and combines with each individual feature.

Experiment	***Feature used for Representation***	***Number of cluster selected***	***Cluster purity (%)using k-means algorithm***	***Cluster purity (%)using EM algorithm***
Fourth	Morphological–position– syntactic – TF	2	80.5	78.5
	Morphological–position– syntactic – IDF		80.54	78.56
	Morphological–position– syntactic–TF*IDF		80.6	78.7

Table 4: Experiment, combined four feature representation, cluster purity using both k-means and EM algorithm

The fourth experiment result shows that k-means value and EM value drop compared to third experiment. Finally, Based on the evaluation of clustering purity result an optimal purity combination of those features is selected by applying forward feature selection method which is taking a mixture of feature which registers high purity value this combination continues until purity drops.

The study proposed to compare K-means and EM algorithm to cluster noun terms in the Amharic text. The study select features to state where the nouns are located in the text. In conducting an experiment the experimental setting is classified into six based on the feature used to cluster. Each data set contains the same feature terms across the experimental setting but with different type of feature. So the result shows that the first dataset contains words with TF value with K-means value 34.3 % and EM algorithm

32.8% cluster purity. The second dataset is with, words and the corresponding IDF value with K-means value ***34.8***% and EM algorithm ***33.2%*** cluster purity. The third one includes TF*IDF value with K-means value ***35.2%*** and EM algorithm ***34.5%*** cluster purity. The fourth one contains positional information in relation to words with K-means value ***51%*** and EM algorithm ***52%*** cluster purity. The next dataset comes with words and syntactic features with K-means value ***41.2%*** and EM algorithm ***40.2%*** cluster purity. The last experiment is done with the sequence of morpheme which is attached to words with K-means value ***72%*** and EM algorithm ***70%*** cluster purity. Based on the first experiment result six different values are registered due to the effectiveness of features to guess category of the word. So the result shows that syntactic features, morpheme and positional information are the better features that enable to category noun terms from other Amharic parts of speech categories.

Each individual value of feature have big impact on the value of remaining results so the second, third and the fourth experiment result are depends on the first experiment. Because of the value of TF, IDF and TF*IDF result drop in each experiment have less impact to cluster noun terms this is due to the size of corpus used because when we use large size of corpus the value TF, IDF and TF*IDF value increase. So the third experiment result shows *Morphological –position- syntactic feature* combination is the final optimal features value compared with other feature combination with k-means value ***80.75 %***and EM value78.9%, which are used for evaluating the performance of the clustering. But the best performance is still less than 100%. Better performance will be achieved by incorporate other features that are not included in this study and if there is a standard large corpus for experimentation.

Evaluation

Cluster analysis or clustering is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense; accordingly this study aims to find out two cluster ,the first word category is noun terms and other part-of –speeches as a result of these shows the confusion matrix. Two algorithms K-means and EM were applied on the data sets.

Table 5 Shows the confusion matrix by K-means algorithm, whereas Table 6 shows that by EM algorithm. The summarized output is shown below

	N (Noun)	Other POS	Total
N (Noun)	2685	444	3129
Other POS	334	609	943
Total	3019	1053	4072

Table 5: Confusion matrix by K-means

This cluster is made from noun and other Part-of-Speeches. This cluster contains 2685 nouns and next common class other part of speeches which is 444. The second cluster contains 609 other Part-of-Speeches and the 334 nouns.

	N (Noun)	Other POS	Total
N (Noun)	2375	503	2878
Other POS	393	801	1194
Total	2768	1304	4072

Table 6: Confusion matrix by EM algorithm

This cluster is made from noun and other Part-of-Speeches. This cluster contains 2375 nouns and next common class other part of speeches which is 503. The second cluster contains 801 other Part-of-Speeches and the 393 nouns. Mis- clustering occurred in both cluster because the feature selected here in this study not enough to separate noun term form the other Part-of-Speeches. According to the performance measure the accuracy by K-means is ***80.75*** percent, whereas the accuracy by the EM algorithm is ***78.9*** percent. By comparing the results, it is found that K-means algorithm works better on the data sets, when compared to EM. Hence for the clustering of data, it is better to apply k-means algorithm.

Conclusion

In this study, an attempt is made to cluster noun terms for Amharic language using unsupervised learning methods. In order to conduct this research various appropriate and related literature resources, such as books, journal articles, conference paper and the internet have been reviewed. This study also used Python programming languages for text pre-processing and WEKA for constructing clustering model. The features used to represent each words are morphological, positional and term weighting such as Term Frequency (TF), Inverse Document Frequency (IDF), Term Frequency *Inverse Document Frequency (TF*IDF) ,Feature are combined using forward feature selection to select the best features. The motive of this study was to compare some of the clustering algorithms in terms of accuracy. Every algorithm has their own significance and we use them on the nature of the data, but on the basis of this study two clustering algorithms considered are K-means and Expectation-Maximization algorithms. According to the performance measure the accuracy by K-means is 80.75 percent, whereas the accuracy by the EM algorithm is 78.9 percent. After the evaluation of the two algorithms, k-means clustering algorithm accuracy performance is better than EM Clustering algorithm.

This research answers the possibility to use unsupervised learning for Amharic noun clustering by comparing two clustering algorithm accuracy and based on the result k-means clustering algorithm is better than EM clustering algorithm. In doing so this research finds the features that are capable of distinguishing noun with other those features are morphology, positional and syntactic information. This study contributes to the reduction of the amount of time and energy spent while developing language-related applications using Amharic noun.

Recommendations

The study identifies future research direction to come up with an efficient unsupervised noun term clustering for Amharic language. Thus, the following are recommended research areas.

1. In this study K-Means clustering and EM clustering are used for implementing noun clustering by comparing accuracy only. Hence we recommend the need to conduct similar researches using both algorithms to evaluate algorithm the time-consuming and other evaluation criteria.
2. One of the limitations of this research is getting large corpus for experimentation. However, to undertake an extensive experimentation there is a need to build

standard large amount of raw corpus. Hence we recommend as future research direction.

3. Since, unsupervised clustering predicts the correct noun term for a word in a given context using unlabeled corpus. Therefore, further research can investigate enhancement of the task of information retrieval for Amharic language.
4. We recommend also conducting similar researches in other local languages by adopting the procedures followed in this study.

REFERENCES

1. Hans van Halteren, Jakub Zavrel, Walter Daelemans (2001). Improving Accuracy in NLP Through Combination of Machine Learning Systems. Computational Linguistics. 27(2): 199–229.

2. Lewis, D. and Croft, W. (1990) Term clustering of syntactic phrases. SIGIR-90, pp. 385-404.

3. Martin, S., Liermann, J., & Ney, H. (1998). Algorithms for bigram and trigram word Clustering. Speech Communication, 24, pp. 19–37.

4. Sarvakar, K. (2014). Density Based Methods to Discover Clusters with Arbitrary shape in weka. International Journal of Research in Information Technology (IJRIT).

5. Christopher D. Manning and Hinrich Schutze. (1999). In Foundation of statistical Natural Languge processing (pp. 495-528). Cambridge, Massachusetts: The MIT press.

6. Donald Hindle.(1990) Noun classification from predicateargument structures. In Proceedings of the 28th Annual Meeting of the ACL,.

7. Bender, M. L., Sydney W. Head, and Roger Cowley .(1976). The Ethiopian Writing System. In Bender et al (Eds.) Language in Ethiopia. London: Oxford University Press.

8. S. Fissaha and J. Haller , Application of corpus-based techniques to Amharic texts, Institute for Applied Information Sciences, University of Saarland

9. Amsalu, S. (2001): The application of information retrieval techniques to Amharic. Master of Science Research work, School of Information Studies for Africa, Addis Ababa University, Addis Ababa,Ethiopia.

10. Alemayehu, N., Willett, P. (2002). Stemming of Amharic words for information retrieval. Literary and Linguistic Computing 17(1), 1–17

11. Argaw, Atelach.A., Asker, L. (2006): Amharic-English information retrieval. (eds.) Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross Language Evaluation Forum, CLEF 2006, Alicante, Spain, Revised Selected Papers, pp. 43–50.

12. Abiyot Bayou. 2000. Developing Automatic Word Parser for Amharic Verbs and Their Derivation, Master Research work at School of Information Studies for Africa, Addis Ababa.

13. Tesfaye Bayou. 2002. Automatic Morphological Analyzer: An Experiment Using Unsupervised and Autosegmental Approach. Masters Research work. Addis Ababa University.

14. Nega.(1999) Development of Stemming Algorithm for Amharic Text Retrieval. PhD research work, University of Sheffield,.

15. Zelalem Sintayehu.(2001).Automatic Classification of Amharic News Items: The Case of Ethiopian News Agency. (Masters Research work). School of Information Studies for Africa. Addis Ababa University. Addis Ababa. (unpublished).

16. Saba Amsalu Tessera.(2001).The Application of Information Retrieval Techniques to Amharic Document s on the Web. (Masters Research work). School of Information Studies for Africa. Addis Ababa University. Addis Ababa. (unpublished).

17. Salton, G., Yang, C. S. (1973). On the Specification of Term Values in Automatic Indexing. Journal of Documentation, 29, 351-372.

18. Sparck Jones, K., Willet, P. (1997) (Ed.) Readings in Information Retrieval. San Francisco, California: Morgan Kaufmann Publishers Inc.

19. N. Sharma , Aman Bajpai , Mr. Ratnesh Litoriya (2012). Comparison the various clustering algorithms of weka tools. Jaypee University of Engg. & Technology

20. Abdur Chowdhury, M. Catherine McCabe, Improving Information Retrieval Systems using Part of Speech Tagging, ISR, Institute for Systems Research,1998-4

21. Taiwo Oladipupo Ayodele ,Types of Machine Learning Algorithms, University of Portsmouth, United Kingdom

22. Madison ,Andrew B. Goldberg(2009) , Introduction to Semi-Supervised Learning, Synresearch work Lectures on Artificial Intelligence and Machine Learning, University of Wisconsin, 130 pages

23. Ying Zhao and George Karypis. (2003). Clustering in Life Sciences. Minneapolis: Humana Press.

24. Dongen, Stijn Marinus van. (2000). Graph clustering by flow simulation. University of Utrecht.

25. P. Behrkin. 2006. Grouping Multidimensional Data Recent Advances in Clustering, chapter A Survey of Clustering Data Mining Techniques, pages 25–71.

26. Springer Berlin Heidelberg ,(1984). ISBN 978-3-540-28349-2.Murtagh.

27. J.O.Ramsey and David Wiley. (1978). Applied Psychological Measurement. West Publishing Co. Karypis, G. (2003). CLUTO a clustering toolkit. Minneapolis.

28. Willett, P. a. (1991). The limitation of term co-occurrence data for query expansion in document retrival system. Journal of the American Society for Information Science , 122-125.

29. Strzalkowski, T. (1994). Robust text processing in automated information retrieval. Proceedings of the Fourth Conference on Applied Natural Language Processing, Stuttgart, Germany: Association for Computational Linguistics. Reprinted in K. Sparck Jones & P. Willet (Eds.).

30. Frank Keller, 2009. Evaluating Models of Syntactic Category Acquisition without Using a Gold Standard. Proc. 31st Annual Conf. of the Cognitive Science Society, 2576–2581.

31. Christodoulopoulos, Christos, Sharon Goldwater, and Mark Steedman,(2010). Two Decades of Unsupervised POS induction: How far have we come? Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics,.

32. A. Singhal, Modern information retrieval,(2001): A brief overview, IEEE Data Engineering Bulletin, vol. 24, no. 3, pp. 35-43,.

33. A. Das and A. Jain(2008), Indexing the World Wide Web: The Journey So Far, Google Inc, USA, vol. 1, no. 1, pp. 1-24.

34. P. Soucy(2005), ―Beyond TFIDF Weighting for Text Categorization in the Vector Space Model, International Joint Conference on Artificial Intellegence, vol. 1, no. 1, pp. 1-6,.

35. Ceglarek D., Haniewicz K. Rutkowski W.(2011), Domain based semantic compression for automatic text comprehension Augmentation and recommendation, Third International Conference, ICCCI 2011, Gdynia, Poland, September 21-23.

36. J. Etzold, A. Brousseau, P. Grimm, and T. Steiner, Context-aware Querying for Multimodal Search Engines, Springer-Verlag Berlin Heidelberg, vol. 6, no. 2012, pp. 728-729, 2011.

37. Y. Fang, N. Somasundaram, L. Si, J. Ko, and A. P. Mathur, Analysis of An Expert Search Query Log Categories and Subject Descriptors, Symposium A Quarterly Journal In Modern Foreign Literatures, vol. 64, no. 18, pp. 1189-1190, 2011.

38. E. Greengrass (2000), Information Retrieval: A survey, Information Retrieval, vol. 163, no. November, pp. 141-163.

39. P. Ingwersen, Information Retrieval Interaction, 1st ed. London: Taylor Graham Publishing, 2002.

40. C. D. Manning, P. Raghavan, and H. Schutze (2009), An Introduction to Information Retrieval, Online Edition, Cambridge: Cambridge UP.

41. Y. Y. Yao(2012), Information Retrieval Support Systems. Boca Raton: Taylor & Francis Group, pp. 1-778,.

42. Brill, Eric (1995) Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. Computational Linguistics.

43. T. Mandl and C. Womser-Hacker(2005). The effect of named entities on effectiveness in crosslanguage information retrieval evaluation. In SAC• f05: Proceedings of the 2005 ACM Symposium on Applied computing, pages 1059.1064, Santa Fe, New Mexico,. ACM Press.

44. T. Wakao, R. Gaizauskas, and K. Humphreys. 1996. Evaluation of an algorithm for the recognition and classification of proper names. In Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), pages 418-423, Copenhagen, Denmark.

45. Adafre, Sisay Fissaha(2005). Part of speech tagging for Amharic using conditional random fields. Proceedings of the ACL workshop on computational approaches to semitic languages. Association for Computational Linguistics.

46. Girma Awgichew, and Mesfin Getachew. Manual annotation of Amharic news items with part-of-speech tags and its challenges.Ethiopian Languages Research Center Working Papers 2 (2006): 1-16.

47. Gamback, B., and Lars Asker (2010). Experiences with developing language processing tools and corpora for Amharic. IST-Africa, 2010. IEEE,.

48. Tachbelie, Martha Yifiru, and Wolfgang Menzel (2009). Amharic Part-of-Speech Tagger for Factored Language Modeling. RANLP.

49. Tachbelie, Martha Yifiru, Solomon Teferra Abate, and Laurent Besacier (2011). Part-of-speechtagging for underresourced and morphologically rich languages—the case of Amharic. HLTD (2011) 50-55.

50. Gebrekidan, Binyam(2009). Part-of-speech Tagging for Amharic. ISMTCL Proceedings, International Review Bulag.

51. Ghassan. K, Riyad, Majdi. S ,Improving Arabic Information Retrieval system using part of speech tagging,university of Yarmouk .

52. Goldwater, Sharon and Thomas L. Griffiths. 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Prague, pages 744–751.

53. Kristina Toutanova and Mark Johnson. (2007). A Bayesian LDA-based model for semi-supervised partof-speech tagging. In Proceedings of NIPS.

54. Zelalem, Unsupervised Part-of-Speech Tagger for Amharic , Addis Ababa University.

55. Mersehazen woldeqirqos (1934) yäamarn säwäsäw, Birhanena Selam Printing Press, Addis Ababa.Getahun Amare. 1998. Zamanawi yaamarEna Sawasaw baqalal aqaararab. Commercial Printing Press: Addis Ababa

56. Dawkins, C.H. 1969 the Fundamentals of Amharic. A.A Sudan interior mission.

57. Baye, Y. (1992). Ethiopian Writing System http://www.ethiopians.com/bayeyima.html

58. Baye Yemam. 1987 ዓ.ም. የአማርኛ፣ሰዋስው። ት.መ.ማ.ማ.ድ. .።

59. Leslau(1995), Wolf. Reference grammar of Amharic. Otto Harrassowitz Verlag.

How to cite this article:
Aklilu Mandefro Messele, “Comparing K-Means and Expectation Maximization Algorithm to Cluster Amharic Noun Terms”, Journal of Multidimensional Research and Review (JMRR), Vol.1, Iss.1, pp. 31-56, 2020