Extractive Based Auto-Summarizer for Amharic News using K-Means Clustering

Jigsa Tesfaye
HiLCoE, Computer Science Programme, Ethiopia
jiks.tes@gmail.com

Wondowssen Mulugeta
HiLCoE, Ethiopia
School of Information Science, Addis Ababa University, Ethiopia
wondwossen.mulugeta@aau.edu.et

Abstract

Today the amount of Amharic digital resource is growing rapidly. A number of news sources like fanabc.com, addisadmas.com and ethopianreporter.com are rolling out news daily. This makes it harder for us to efficiently read and extract relevant information from all these sources. On the bright side, the availability of these machine-readable Amharic news content creates the opportunity to easily accumulate statistical learning that could reveal important language patterns to develop unsupervised machine learning algorithms for the language. This paper mainly focuses on the design of an extractive automatic summarizer which uses statistical knowledge to intelligently determine the optimum extraction level and content using K-means unsupervised algorithm and it is tested specifically on Amharic news content. This approach was previously used in a similar way for an English language summarizer, but in this paper an improved selection process is provided which combines the restrictive feature of the conventionally known best-n method with the K-mean. Best-n is a method used to have control over the level of percentage of content extracted to produce the final summary. An evaluation was conducted to compare the two ways of sentence selection; the improved K-means method with the conventional best-*n*%. The K-means method showed a decrease in extraction percentage with an improved summary score of 58.1% which is a 5% increase from the conventional best *n*% selection method.

Keywords: Automatic Summarization; K-means; TF-IDF; Web Crawler; News Content

1. Introduction

Text summarization is the process of producing a shortened version of a given text retaining the most relevant information for the reader. Summarization can lessen the burden of information processing by condensing the original text into one which is more or less representative.

In this paper, an improved approach is introduced that uses K-means algorithm for selection of content in a language independent extractive summarizer. In previous researches, including such kind of language independent implementation researches done on Amharic, the prototype summarizers were designed to produce a summary with a strict level of extraction in percentage (usually 20% and 30%) or best *n*%; with *n* decided by the user. The best *n* approach ranks sentences based on their scores and selects the top *n* sentences. The problem with this approach is that when ranking the sentences, the score gap could be very critical for the selection process. Let us say we have twenty sentences in a document and the 6th and 7th ranked sentences have huge score gap and if the user assigns *n* as 35% then the inclusion of the last sentence has big negative influence on the accuracy of the summary.

K-means is a clustering algorithm that is used to model unsupervised data. In K-means, data objects are presented as points on a Cartesian plane. Starting with a random set of K reference points on n-dimensional space, the data points will be assigned to K clusters based on distance criterion. Centroids are special points that represent a cluster. These centroid points are selected with a calculated process that attempts to choose the points placed as far away from

each other as possible [1], setting as much distance among clusters as possible. The time spent when finding these ideal centroids is divided into iterations. The steps for the K-means algorithm are as follows:

1. On a single or multi-dimensional plane, the objects are positioned as points based on their features as a metric.
2. K points are selected from the objects as centers of clusters, also known as centroids.
3. Every data sample is assigned to a cluster with the closest centroid.
4. The mean of the data points is calculated for each cluster possibly choosing a new more representative centroid in the process.

The third and fourth steps are iterated until no changes happen in the centers of clusters. K is the number of clusters and should be defined ahead of time.

The content scoring algorithm used in this paper is Term frequency, inverse document frequency (TF-IDF). TF-IDF is a statistical method used to compute the significance of a word in a document by looking at its occurrence frequency in the document compared to a set of other similar documents [2]. Given the term ‘t’ in a document 'd', the TF value of ‘t’ is calculated as the frequency of ‘t’ in ‘d’ over the total number of all the terms found in ‘d’ as:

$$\mathbf{tf}(t,d) = f_{\downarrow}(t,d) / \Sigma_{\top}(t^{\intercal\prime} \in d)\, f_{\downarrow}(t^{\intercal\prime}, d)$$

The IDF of a word is assigned the inverse value of the number of documents it is found normalized by the total number of documents as:

$$\mathbf{idf}(t,D) = \frac{\log N}{|\{d \in D : t \in d\}|}$$

TF-IDF is a numeric weight given to a word which is a product of two formulas, TF and IDF.

$$\mathbf{tfidf}(t,d,D) = \mathbf{tf}(t,d) \cdot \mathbf{idf}(t,D)$$

The representation of sentences in the k-means clusters can be set with a numerical score value of an attribute or combined attributes [2]. Indicators such as term frequency and inverse document frequency can represent a relevance factor of a sentence based on frequency.

2. Related Work

Even though there weren’t as such similar summarization researches that specifically use TF-IDF and clustering on Amharic language, there are quite many papers that share most parts of the methodologies and algorithmic techniques with this paper [3, 4, 5]. The use of news articles and newspapers for developing and testing a summarizer is very common in the research community [6, 7]. The ease of finding large collection of news digital documents is a compelling reason that made researchers so interested to study the news domain.

The work in [1] is the closest in terms of the algorithms used and overall methodology. They used TF-IDF algorithm as a feature and K-means clustering for summarization. Sentences are scored using words’ TF-IDF sum to measure “goodness”. The difference in their approach comes down only to two decisions; the way k is chosen dynamically and the method used in cluster selection. Each sentence is represented as coordinates in a single dimensional Cartesian plane which will then be clustered using K-means. When selecting the sentences, they argue that the cluster with the maximum number of sentences should be selected. K, which represents the number of clusters to be created, is dynamically determined with the possibility of having k value more than two (as shown below) that was designed after running tests on multiple documents that calculate dynamic k.

If N>20 then k= N-4

Else k= N-20

where N is the total number of sentences in the document.

The authors argue that this algorithm will balance the amount of extraction with the size of the document. As implemented in this paper, their approach cluster sentences and in the end one cluster is chosen as a summary. The sentences are presented in the summary in their order found in the original document. But the work also had a limitation in a

way it produced an average of 43% extraction with the rate going up to 50% which only begs the question if 50% extraction really can be considered a summary.

This K-means and TF-IDF implementation is language independent. Past research implementation of Amharic summarizers incline towards a language independent, single document, unsupervised and extractive based approaches. There hasn't been as such many works done on the language with handful of researches only limited to the use of basic content indicators, topic modeling algorithm and Latent semantic analysis techniques. Almost all works used the best-n approach with different fixed extraction rates – 20%, 30% in [8], 30%, 25%, 20% in [9] and 100 words in [10].

3. The Proposed Solution

3.1. Prototype Design

The prototype is a program that has an intuitive user interface that takes input of an Amharic news article content in plain text format and a preferred level of content extraction in percentage to produce a summary output. The final summary consists of actual extracts of sentences that are deemed to have special relation and importance with the central topic of the writing. The summarizer is an unsupervised system that considers keywords or content words as summary indicators and uses frequency driven technique for scoring.

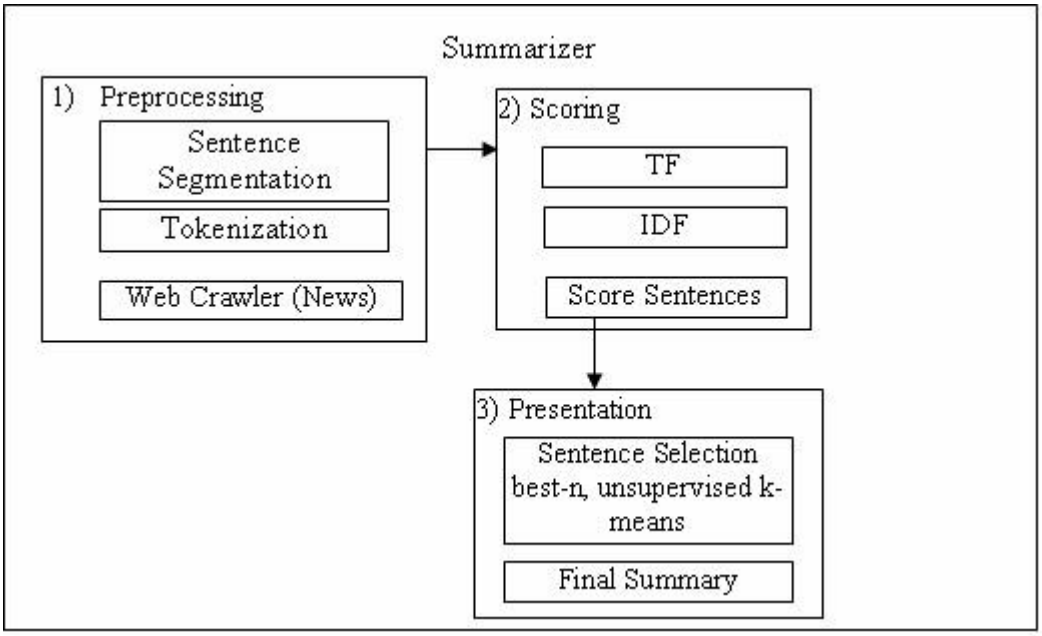

Figure 1: The Architecture of the Prototype System

The summary extraction process programmed in the prototype has distinct stages from preprocessing, intermediate representation, scoring, up to selection and finally summary presentation. The most important stage is scoring because it guides all the other stages. The scoring is done using frequency driven topic terms identification algorithm TF-IDF. The final stage of summary selection can be programmed to be determined using two options, the K-means or pure best *n* methods. Clustering is an unsupervised technique used in this particular case to select the most outstanding sentences from the input document. K-means is the clustering algorithm used to achieve this based on the TF-IDF score representation of the sentences. As mentioned before, this paper aims to evaluate an alternative K-means implementation approach for determining the summary extraction percentage threshold level automatically by identifying the outstanding cluster of sentences using their centroid distance measure. In a previous work in [1] the level of percentage of extraction was not limited when using a pure K-

means selection. In our case the K-means is processed in increased iteration depending on the extraction level set by the user which makes it a hybrid approach with the Best-n. The prototype should also support the conventional sentence selection method in the form of the best *n*% where the user specifies a rigid percentage level in which case the top ranked sentences are automatically added in the summary. These two selection methods are discussed more in Section 3.3. The architecture of the prototype system is shown in Figure 1.

3.2. Data Collection

The prototype is tested on Amharic language specifically the news domain. Amharic is still under resourced that hindered the development of language processing applications. Consequently, it is necessary to explore ways to efficiently populate resources. For languages like Amharic where most of the resources available are in unstructured text, in the form of news articles and messages, the design of unsupervised systems could be the way forward at the moment [11].

Collection of large amount of online news data can be very useful for identifying keywords that can in turn be used to produce a short summary of news articles [12]. The main content scoring algorithm used in this research is TF-IDF. In order to calculate the IDF of words there needs to be a considerably large amount of corpus. A document frequency is the number of documents a term is found in a corpus. The collection of documents from which the word document frequency knowledge is gathered should be similar to the document on which the system algorithm applies, specially for domain specific tasks like this [13]. Therefore, the corpus data is gathered from a large collection of news articles. There are commonly found terms in news articles such as "ዘገባ", "ቃለመጠይቅ", "ገለጹ". The use of this news corpus allows us to identify such words that are often used in the news domain. The scores of these terms should be lowered depending on their IDF value. IDF was designed for the purpose of achieving a frequency driven weighting score for terms. As its name suggests IDF holds the inverse value of the number of documents a term was found to appear in.

An automated web crawler program developed specifically for this research, adopted from utopiaio'saScraper implementation (found at github.com/utopiaio/eKeyboard), was used to help build the news corpus. This crawler automatically goes through hundreds of news web pages on the Internet to grab contents of news articles. So the document frequency information of each appearing word is then collected and shared to the summarizer application.

3.3. Sentence Selection Stage

The implementation of the K-means from this paper's perspective begins with a bag of words approach [14] where each word from each sentence is gathered and viewed collectively. The sentences are then represented by their scores in a single dimensional Cartesian plane where score is calculated as:

$$\text{Score}(S) = \sum_t \text{TF-IDF}_t / |S|$$

where S is sentence, |S| is sentence length and t are terms in the sentence.

These coordinates are used as input for unsupervised K-means clustering algorithm. As pointed out in [15], one of the proposed methods is to take points whose attribute values are average of the n-objects from the dataset. Therefore, the initial points are the 'k' sentences that are found in the middle of the score ranking. Finally, the cluster with the highest centroid value is chosen as the appropriate summary. The rationale behind this centroid-based summarization is that the centroid has a mean value representing the cluster. Therefore, a centroid with higher value signifies the sentences in the cluster have bigger relevance.

Algorithm 1: Best-N

```
Input: Percentage of extraction
  preferred by the user, list of
  sentences and their scores
Begin
  Rank the sentences based on their
    score
```

```
    Take the best n% sentences and
      present them to the user in the
      order they were presented in the
      original document
End
```

Algorithm 2: K-Means

```
Input: Maximum percentage of
  extraction preferred by the user,
  list of sentences and their scores
Begin
  Set k initial value to 2
  Rank the sentences
  While (percentage of selected
    sentences > restriction set by
    the user & k <= 10) {
    Choose the most average k
      sentences as initial points
      based on their rankings
    Cluster the sentences using k-
      means
    Choose the cluster with the
      highest centroid value as a
      possible summary
    k++
  }
  Present the sentences to the user
    in the order they were presented
    in the original document
End
```

There are two types of extractions, K-means and pure Best-n%. If the selection type is Best-n% (see Algorithm 1), the user selects percentage level option from 5 to 95% and the top n% of sentences directly qualify to the summary. If it is set as K-means selection (see Algorithm 2), then the sentences first get clustered based on their scores and the outstanding cluster of sentences with high centroid get selected. The user can set a restriction as the maximum level of extraction for the K-means. The K value depends on the restriction set by the user, for example, the user could set the restriction at 30% or less.

4. Discussion

For the evaluation, this paper presented results from experiments done to test the accuracy of the prototype from two perspectives. The first one tests the precision and recall measures of the conventional Best-n implementation. The second one checks summaries produced by K-means where clustering technique is used to determine varying level of extraction with a restriction of 30%. The results show a better F-measure score for the K-means by 5%. This is because the K-means reduces the extraction level by taking out irrelevant sentences from the summary increasing the recall score. The average extraction rate decreased by 6.75% when using K-means which is good because it is a shorter summary with better accuracy. But the precision is lowered by about 5%. This was because the best-n summaries taken were allowed higher percentage extraction. The system overall retained a 58.1% F-measure score from an evaluation done on four articles by comparing summaries received from four people.

When compared to other researches on Amharic language, the document set used for testing are long with an average length of 40 sentences and 1075 words and also had less extraction percentage average of 18%. The alternative K-means approach presented in this paper tries to balance the extraction level better when compared with the approach taken in [1] where they tested pure K-means for extraction and had summary results more than 40% in most cases, average of 43.2% and a maximum of 50%. The test cases in this evaluation show average of 18%, 25% lower. To support the content scoring, the crawler was successfully used to mine frequency data of 145,650 words from three well-known news sources.

5. Conclusion

This main focus was on the design of selection process for auto summarizers. It was attempted to provide an optional and improved implementation of the K-means as a selection method judging from a previous research in [1]. It showed how clustering algorithms like K-means can be integrated with TF-IDF to produce a more unsupervised summary. The experiments in this paper showed an approach that integrates K-means to produce a restrictive summary less than or equal to 30%. K-means is applied iteratively to determine k (number of clusters) and

take the highest centroid valued cluster as a summary. We conclude that this method improves the accuracy of the best-n by also solving the high percentage summary problem observed in [1].

The experiments focused on the comparison of the two summary selection approaches in the form of the newly suggested K-means unsupervised selection and the conventionally known best-n method. The results of the experiments show 5% increase in accuracy with the K-means cluster based selection method. The new approach described in this paper showed a more appropriate level of extraction unlike that in [1] with the average extraction level of 18%.

To counter the scarcity of resource in Amharic language this research developed a dedicated web crawler application and presented an approach on how to integrate this application as a subordinate service to compliment the summarizer and ensure a scalable and adaptable approach. This in turn showcased a method on how a useful corpus can be prepared automatically from Amharic news websites.

References

[1] Agrawal, A. and Gupta, U., "Extraction Based Approach for Text Summarization Using K-Means Clustering", International Journal of Scientific and Research Publications, 4 (11), 2014.

[2] Munot, N. and Govilkar, S. S., "Comparative Study of Text summarization Methods", International Journal of Computer Applications, 102 (12), 2014.

[3] Seki, Y., "Sentence Extraction by TF-IDF and Position Weighting from Newspaper Articles", The Third NTCIR Workshop, National Institute of Informatics, 2002.

[4] Christian, H., Agus, M. P., and Suhartono, D., "Single Document Automatic Text Summarization Using Term Frequency-Inverse Document Frequency (TF-IDF)", 2016.

[5] Radev, D. R., Fan, W., and Zhang, Z., "WebInEssence: A Personalized Web-Based Multi-Document Summarization and Recommendation System", In NAACL Workshop on Automatic Summarization, Pittsburg, 2001.

[6] Jassem, K. and Pawluczuk, Ł., "Automatic Summarization of Polish News Articles by Sentence Selection", In proceedings of the Federated Conference on Computer Science and Information Systems, 2015, pp. 337-341.

[7] Zadbuke, A., Pimenta, S., Padwal, D. and Wangikar, V., "Automatic Summarization of News Articles using TextRank", Special Issue on 3rd International Conference on Electronics & Computing Technologies, 2016, pp. 124 - 127.

[8] Melese Tamiru, "Automatic Amharic Text Summarization Using Latent Semantic Analysis", Unpublished Masters Thesis, Addis Ababa University, 2009.

[9] Eyob Delele, "Topic-based Amharic Text Summarization", Unpublished Masters Thesis, Addis Ababa University, 2011.

[10] Mattias, Gesesse Argaw, "Efficient Language Independent Text Summarization Using Graph Based Approach", Unpublished Masters Thesis, Addis Ababa University, 2015.

[11] Bekele, Worku Agajyelew, "Information Extraction from Amharic language Text: Knowledge-poor Approach", Unpublished Masters Thesis, Addis Ababa University, 2015.

[12] Gupta, V. and Lehal, G. S., "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, 1 (1), pp. 60-76, 2009.

[13] Shams, R. and Elsayed, A., "A Corpus-based Evaluation of Lexical Components of a Domain-Specific Text to Knowledge Mapping Prototype", IEEE 11th International Conference on Computer and Information Technology, 2008.

[14] Gupta, V. and Lehal, G. S., "A Survey of Text Summarization Extractive Techniques", Journal of Emerging Technologies in Web Intelligence, 2 (3), 2010, pp. 258-267.

[15] Baswade, A. M. and Nalwade, P. S., "Selection of Initial Centroids for K-Means Algorithm", International Journal of Computer Science and Mobile Computing, 2 (7), 2013, pp. 161-164.