Indian Journal of Science and Technology, Vol 10(17), DOI: 10.17485/ijst/2017/v10i17/106493, May 2017
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Evaluation of Unsupervised Learning based Extractive Text Summarization Technique for Large Scale Review and Feedback Data

Jai Prakash Verma[1]* and Atul Patel[2]

[1]Department of CE, Institute of Technology, Nirma University, Ahmedabad – 382481, Gujarat, India; jaiprakash.verma@nirmauni.ac.in
[2]CMPICA, CHARUSAT University, Changa – 388421, Gujarat, India; atulpatel.mca@charusat.ac.in

Abstract

Background/Objectives: Supervised techniques uses human generated summary to select features and parameter for summarization. The main problem in this approach is reliability of summary based on human generated parameters and features. Many researches have shown the conflicts in summary generated. Due to diversity of large scale datasets, supervised techniques based summarization also fails to meet the requirements. Big data analytics for text dataset also recommends unsupervised techniques than supervised techniques. Unsupervised techniques based summarization systems finds representative sentences from large amount of text dataset. **Methods/Statistical Analysis:** Co-selection based evaluation measure is applied for evaluating the proposed research work. The value of recall, precision, f-measure and similarity measure are determined for concluding the research outcome for the respective objective. **Findings:** The algorithms like KMeans, MiniBatchKMeans, and Graph based summarization techniques are discussed with all technical details. The results achieved by applying Graph Based Text Summarization techniques with large scale review and feedback data found improvement over previously published results based on sentence scoring using TF and TF-IDF. Graph based sentence scoring method is much efficient than other unsupervised learning techniques applied for extractive text summarization. **Application/Improvements:** The execution of graph based algorithm with Spark's Graph X programming environment will secure execution time for this types of large scale review and feedback dataset which is considered under Big Data Problem.

Keywords: Extractive Text Summarization, Sentence Scoring Methods, Unsupervised Learning

1. Introduction

Content based Recommendation system for large amount of text data generated by different stake-holders for an organization in the form of review and feedback. These types of text data are generated from different types of computerized automated feedback and review system or extracted from web. This is a type of Big Data Analytics problem because of data volume, velocity and variety[1]. An extractive text summarization based recommendation system model is proposed for analyzing and summarizing these large amount of text data known as Big Data. The system helps finding actionable insights for better decision making. Such Text Summarization Systems can be categorized as per following categories.

Extractive and Abstractive Summarization Systems[2–4] are the techniques to summarize large amount of text using computer programs. Extractive summarization technique is based on selection of representative text from the given large text data. Abstractive summarization is generating summary, based on the sense and feeling of the text document. Here in abstractive summary, we may use new words for sensing the large text but in extractive summarization program will identify representative set of words and sentences. Single and Multi-Document Summarization Systems[5,6] categorize text summariza-

*Author for correspondence

tion systems based on approach in which number of documents are selected for analyzing the dataset. Multi document summarization is more complex than single document summarization but it recommends for review and feedback summarization. Generic and Query-Based Summarization Systems[7–9]categories summarization techniques between a specific request based and generic query based summarization. Generic summary is based on main topics covered in the text but query based summarization specifies the request or question for summary. Another categorization of text summarization systems are based on the Supervised and Unsupervised Techniques[10–12]. Supervised techniques use dataset which are annotated by humans before applying the algorithm but unsupervised techniques do not use this type of human annotations with dataset. Unsupervised techniques use the linguistic and statistical information generated from the dataset for text summarization. Another categorization in text summarization systems are based on Surface-Level and Deeper-Level Summarization Systems[13]. Surface-Level and Deeper-Level Summarization Systems summarize the text as per the purpose of summary. Generally this type of summarizations are used for news articles, scientific text etc.

Extractive Text Summarization selects representative sentences from available large scale text dataset. These sentences are selected based on different methods. One method is based on Surface Level Approaches[2,3] in which sentences are selected based on the most frequent words. This type of method gives good results for query and purpose based summarizations but summarization for reviews and feedback is not appropriate for it. Another method is based on Statistical Approaches[2,14] which gives the summary based on relevance of information extracted from dictionaries. For finding relevance information about the selected text classifier, algorithms like Bayesian classifier are used. Another type of method is based on Text Connectivity Approaches[2,4]. In this approach, text summarization is generated by the connectivity of sentences and text based on lexical chains and Rhetorical Structure. Another type of method is based on Graph Based Approaches[4]. The nodes of directed graph represent sentences of text, and edges represent the similarity between these sentences. Summary is generated by selection of sentences with highest similarity associated. Another method is based on Machine Learning Based Approaches[13]. The machine learning based summarization algorithms use techniques like Naïve-Bayes, Decision Trees, Hidden Markov Model, Log-linear Models, and Neural Networks. Algebraic Approaches[2–4] such as Latent Semantic Analysis (LSA), Non-negative Matrix Factorization (NMF), and Semi-discrete Matrix Decomposition (SDD) are also used for text summarization.

Remaining sections of the paper comprise as follows. In section 2 sentence scoring based text summarization techniques are discussed. In section 3 unsupervised learning based text summarization techniques are discussed. In section 4 evaluation methodologies for text summary generated are discussed. Section 5 produces the proposed research work, its techniques and approaches. In section 6, experimental study is described for proposed research work. Section 7 discusses performance analysis based on evaluation methodologies. In Section 8, we present conclusion and future extension in the research work proposed.

2. Sentence Scoring Based Text Summarization

Sentence scoring methods discussed in many research papers basically emphasize on word score, sentence score and graphs, where word score and sentence scores are counted based on the frequencies of word in given text dataset. The graph based sentence scoring is based on relationship between the sentences. The focus of many researches is on analysis of large scale text available or written with print media. Research work in this paper is focused on analyzing the large amount data extracted from web in the form of review and feedback about an enterprise or organization for their products and services.

3. Unsupervised Learning Based Text Summarization

Supervised techniques use human generated summary to select features and parameters for summarization. The main problem in this approach is reliability of summary based on human generated parameters and features. Many researches have shown the conflicts in summary generated. Due to diversity of large scale datasets, supervised techniques based summarization are also not fitted. Study and research on big data analytics for text dataset also recommends unsupervised techniques and their acceptance than supervised techniques[15–17]. Unsupervised

techniques based summarization systems find representative sentences from large amount of text dataset.

4. Evaluation Methodology

Evaluation methodologies for summary generated by different techniques are mainly comparing computer generated summary with human made summary. Here few methods are discussed for evaluating computer generated summary based on their effectiveness and usability. Text Quality Evaluation: The text should not contain any grammatical error such as incorrect words or punctuation errors. Co-Selection Evaluation: Where extracted summaries are compared with ideal summaries.

Content-Based Evaluation: Compare extracted and ideal summaries, even though they do not share sentence. For content-based evaluations, measures such as cosine similarity, longest common subsequence, pyramids, and ROUGE scores are used. Task-Based Evaluation: Compared according to their performance of accomplishing the given task.

4.1 Co-Selection Evaluation

Co-Selection Text Summary Evaluation technique is based on comparison of gold summary with computer generated summary. The main metrics for this method are recall, precision, and f-measure. Recall is the number of terms in both the summaries divided by total number of terms in gold summary (relevant terms) (equation 1). Precision is the total number of terms in both the summaries divided by number of terms in computer generated summary (equation 2). F-measure is a composite measure of recall and precision. Due to contradiction between recall and precision measure, researchers recommend harmonic average of both the measures for performance evaluation as f-measure (equation 3).

$$\text{Recall} = \frac{|\{relevant\ terms\} \cap \{retrieved\ terms\}|}{|\{relevant\ terms\}|} \quad (1)$$

$$\text{Precision} = \frac{|\{relevant\ terms\} \cap \{retrieved\ terms\}|}{|\{retrieved\ terms\}|} \quad (2)$$

$$\text{F-Measure} = \frac{2 \times Precision \times Recall}{Recall + Precision} \quad (3)$$

4.2 Content-Based Evaluation

Co-selection based evaluation measures the number of match terms or sentences between both the summaries. It ignores the fact that both the terms and sentences contain the same information even though both are written differently. Content- Based Evaluation measure can overcome these issues. Cosine Similarity is a basic content-based evaluation measure (equation-4), where X and Y represent sentences or terms in both the summaries.

$$\text{CosineSim}(X, Y) = \frac{\sum_i xi \times yi}{\sqrt{\sum_i ([xi)]^2} \times \sqrt{\sum_i ([yi)]^2}} \quad (4)$$

5. Proposed Research Work

Research work in the area of extractive text summarization with unsupervised learning approach is proposed in this paper. An experimental analysis of unsupervised techniques with python programming language is implemented and discussed. All the steps like data extraction, cleaning, preprocessing, and analyzing for text summary generation are discussed and implemented. As per Figure 1 proposed work highlights the expected data sources, preprocessing steps, and analysis processes. Due to large amount of data, this work also recommends open source solutions for handling the data in effectively and efficiently.

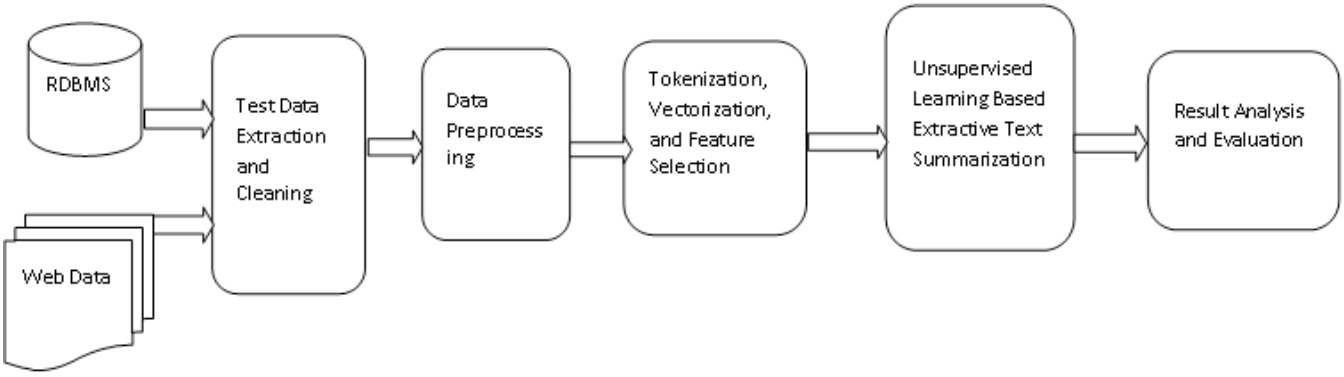

Figure 1. Process flow for proposed research work.

6. Experimental Analysis

In this work, a recommendation system model is proposed to find actionable insights from the text data extracted from different web pages and different computer automated review or feedback systems.

The steps for the proposed model are

6.1 Dataset Selection

Following dataset was used in this research work for experimental study.

Opinion Dataset 1.0[4]: This dataset contains sentences extracted from reviews on a given topic. The Opinion dataset also comes with human composed summaries for evaluation. Feedback and Review data for an educational institution.

Primary Source: Different feedback systems.

Secondary Source: Data Extracted from Facebook and website. Amazon's Review and Feedback data for different products (Size: 34 GB).

6.2 Data Preprocessing

Following are the data pre-processing steps selected for this research work.

1) Remove Punctuation characters
2) Lowercase Conversion
3) Remove numerals
4) Spelling Corrections
5) Singularization
6) Converting all words in Base Form
7) Stop-words Removal.

Data Preprocessing Steps Execution for presentation: As shown in Figure 2 one text file demoTest.txt is used as Input text (python program implementation).

6.3 Tokenization and Vectorization

Tokenization is a process for breaking a text stream in words, sentences, phrases and other meaningful objects called tokens. Vectorization converts these text documents to matrix of token based on their occurrences. Following are the methods (implemented in python) used for both the task:

```
hasher = TfidfVectorizer(input = dataset, max_df=0.5, min_df=2, stop_words='english',use_idf=1)
vectorizer = make_pipeline(hasher, TfidfTransformer())
X = vectorizer.fit_transform(dataset)
```

Data Preprocessing Steps for Extractive Text Summarization		
Data Preprocessing Step	**Example Input**	**Output**
Remove Punctuation characters	Hello world, testing texts for 6 reviews @ Charusat Univorsity	Hello world testing texts for 6 reviews Charusat Univorsity
Lowercase Conversion	Hello world testing texts for 6 reviews Charusat Univorsity	hello world testing texts for 6 reviews charusat univorsity
Remove numerals	hello world testing texts for 6 reviews charusat university	hello world testing texts for reviews charusat univorsity
Spelling Corrections	hello world testing texts for reviews charusat university	hello world testing texts for reviews None university
Singularization	hello world testing texts for reviews None university	hello world testing text for review None university
Converting all words in Base Form	hello world testing text for review None university	hello world test text for review None university
Stop-words Removal	hello world test text for review None university	hello world test text review university

Figure 2. Data preprocessing steps for extractive text summarization.

Feature Selection: In machine learning and statistics, feature selection is the process of selecting a subset of terms from the large amount of terms generated by tokenization and vectorization step for use in model construction.

Cluster and Summary Analysis: Three approaches are selected for Summary Analysis.

a) Sentence Score based on word TF.
b) Sentence Score using Word Base Form' TF.
c) Sentence Score using Word Base Form' TF-IDF.

Following are the python implementation code that have used for clustering analysis. MiniBatchKMeans is modified version of KMeans algorithm which is more efficient for large scale web data. Due to large size and time constraint for cluster MiniBatchKMeans perform more effectively compare to KMeans. Suppose we have a dataset of 500000 reviews and feedbacks, and objective is to divide them into 100 clusters. The complexity of the original K-Means clustering algorithm is O(n*K*I*f), where n is the number of records, K is the number of clusters, I is the number of iterations and f is the number of features. It can be clearly seen that this will take a lifetime for the original algorithm to cluster data. In this research work review or feedback given by individual is considered a document. Due to large size the small size subsets are selected form original dataset and then apply the algorithms for clustering. The algorithm takes small batches (randomly chosen) of the dataset for each iteration. It then assigns a cluster to each data point in the batch, depending on the previous locations of the cluster centroids.

For Kmeans:

km = KMeans(n_clusters=8, init='k-means++', max_iter=100, n_init=1,verbose=0)

For MiniBatchKMeans:

km = MiniBatchKMeans(n_clusters=8, init='k-means++', n_init=1,init_size=1000, batch_size=1000, verbose=0)

Graph Based Text Summarization: Python program is implemented for graph based text summarization where sentences are represented by vertices and similarity between sentences are represented by edges between vertices. It is an unsupervised learning based approach for extractive text summarization by automatic sentence extraction using graph based ranking algorithms. The results achieved by applying Graph Based Text Summarization techniques with large scale review and feedback data found improvement with previously published results based on sentence scoring using TF and TF-IDF. In short, a graph-based ranking algorithm is a way of deciding on the importance of a vertex within a graph, in this research work the vertexes represent the review or feedback given by individual, by taking into account global information recursively computed from the entire graph, rather than relying only on local vertex-specific information. Participating review and feedback in summary sentences are well connected to other sentences. The connectivity of the sentences which is represented by vertexes is based on similarity with other sentences. Similarity measure like TF-IDF can be selected as per performance of the system. Graph G (V, E), where V: set of sentences and E: similarity between sentences. A threshold value is decided for similarity between the sentences. Sentence score is calculated based on the Rank of sentences which is estimated by their degree. Top k sentences are selected for summarizing sentences.

7. Performance Evaluation

As per Table 1, the outcome from above experimental analysis shows that unsupervised techniques for extrac-

Table 1. Comparison of Unsupervised extractive text summarization techniques

Algorithm ID	Average_Precision	Average_Recall	Average_f-measure	Average_Similarity
Sentence Score Based on word TF	0.137966126452	0.206132266328	0.164604495226	0.224801884677
Sentence Score using Word Base Form' TF	0.140276737872	0.309796632534	0.191952080092	0.239164737438
Sentence Score using Word Base Form' TF-IDF	0.123298003326	0.266903172868	0.167818834179	0.220944272061
Kmeans	0.147341297188891	0.324066229588176	0.201706815726618	0.153989154237332
MiniBatchKMeans	0.151333386161	0.331919322357	0.207004641971	0.144447037715
GraphBased Summarization	0.155983580672	0.339218973418	0.212496582933	0.375529302366

tive text summarization improve the recall, precision, and f-measure. MiniBatchKMeans improves the result than K-Means. Graph Based Text Summarization improves the results with recall, precision, and f-measure. Here we are comparing unsupervised learning techniques with sentence scoring methods for extractive text summarization.

8. Conclusion

An unsupervised learning based extractive text summarization system is implemented and evaluated with different algorithms. Graph based sentence scoring method is implemented and evaluated with traditional sentence scoring methods. Programming with Spark programming framework on Hadoop Distributed File System storage is better for efficient execution when compared to other Map Reduce with Hadoop environment. Graph based sentence scoring method gives comparatively better result than other unsupervised learning techniques applied for extractive text summarization. Analyzing Amazon's Review and feedback dataset can provide the future enhancement in this work.

9. References

1. Verma JP, Patel B, Patel A. Big Data Analysis. Recommendation System with Hadoop Framework, IEEE International Conference on Computational Intelligence & Communication Technology. 2015. p. 1–6. PMCid:PMC4410521
2. Ferreira R, Cabral LS, Lins RD, Silva GP, Freitas F, George DC, Cavalcanti A, Lima RA, Steven J, Simske B, Favaro L. Assessing sentence scoring techniques for extractive text summarization. Expert Systems with Applications. 2013; 40:5755–64. Crossref
3. Xiang Z, Schwartz Z, John H, Gerdes J, Uysal M. What can big data and text analytics tell us about hotel guest experience and satisfaction. International Journal of Hospitality Management. 2015; 44:120–30. Crossref
4. Ganesan K, Zhai C, Han. Opinosis. A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions. Proceedings of the 23rd International Conference on Computational Linguistic, Beijing, China: 2010. p. 1–9.
5. Ittoo A. Text analytics in industry: Challenges, desiderata and trends. Comput Industry. 2016. Crossref.
6. Khan A, Salim N, Obasa AI. An Optimized Semantic Technique for Multi- Document Abstractive Summarization. Indian Journal of Science and Technology. 2015 Nov; 8(32):1–11. Crossref
7. Lloret E, Palomar M. Tackling redundancy in text summarization through different levels of language analysis. Computer Standards & Interfaces. 2013; 35:507–18.
8. Bridge D, Healy P. The GhostWriter-2.0 Case-Based Reasoning system for making content suggestions to the authors of product reviews. Knowledge-Based Systems. 2012; 29:93–103. Crossref
9. Online Shopping touched new heights in India in 2012. Hindustan Times, 31 December 2012. 2014 July; 3(7):1–7, Retrieved on 31 December 2012.
10. Bing LI, Keith CC, Chan. A Fuzzy Logic Approach for Opinion Mining on Large Scale Twitter Data. IEEE/ ACM 7th International Conference on Utility and Cloud Computing, 2014. p. 652–7.
11. Ghorpade T, Ragha L. Hotel Reviews using NLP and Bayesian Classification. International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai: 2012 Oct 19-20; 84(6):17–22.
12. Khan A, Baharudin B. Sentiment Classification Using Sentence-level Semantic Orientation of Opinion Terms from Blogs IEEE. IEEE. 2011; 1–17.
13. Thiago S, Guzella, Walmir M, Caminhas. A review of machine learning approaches to Spam filtering. Elsevier Journal - Expert Systems with Applications. 2009; 36:10206–22. Crossref
14. Sheshasaayee A, Jayanthi R. A Text Mining Approach to Extract Opinions from Unstructured Text. Indian Journal of Science and Technology. 2015 Dec; 8(36):1–4. Crossref
15. Nomoto T, Matsumoto Y. A New Approach to Unsupervised Text Summarization. SIGIR'01, Septe, New Orleans, Louisiana, USA: 2001. p. 1–9.
16. Sulthana AR, Subburaj R. An Improvised Ontology based K-Means Clustering Approach for Classification of Customer Reviews, Indian Journal of Science and Technology. 2016 Apr; 9(15):1–6. Crossref
17. Anuradha G, Varma DJ. Fuzzy Based Summarization of Product Reviews for Better Analysis. Indian Journal of Science and Technology. 2016 Aug; 9(31):1–9. Crossref