EasyChair Preprint

№ 2426

Text Summarization Framework Using Machine Learning

Pallavi Kohakade and Sujata Jadhav

January 20, 2020

Text Summarization Framework Using Machine Learning

Mrs.Kohakade Pallavi S.
PG Student,Computer Engineering
Vishwabharati Academy College of Engineering
Savitribai Phule Pune University
Ahmednagar, India
Kohakadepallavi1893@gmail.com

Prof.Jadhav Sujata A.
Department of Computer Engineering
Vishwabhartis Academy College of Engineering
Savitribai Phule Pune University
Ahmednagar, India
jadhav.suj14@gmail.com

***Abstract*—Automatic text summarization is an essential natural language processing application that goals to summarize a given textual content into a shorter model. The fast growth in media information transmission over the Internet demands text summarization using neural network from asynchronous combination of text. This paper represents a framework that utilizes the techniques of NLP technique to examine the elaborative information contained in multi-modal statistics and to enhance the aspects of text summarization. The basic concept is to bridge the semantic gaps among text content. After, the generated summary for important information through multi-modal topic modeling. Finally, all the multi-modal factors are considered to generate a textual summary by maximizing the importance, non-redundancy, credibility and scope through the allocated accumulation of submodular features. The experimental result shows that Text Summarization framework outperforms other competitive techniques.**

***Index Terms*—Summarization,Feature selection,Machine Learning,Sentence Embedding**

I. Introduction

Now a days, there are large numbers of documents or information that is present related to any particular field[1][3]. There are many sources out of which we can gather a lot of information that will be pertinent to our field of search. Much information is available at various sources like the internet. But, as we know that a huge amount of information cannot be always considered or taken into use. So, a precise amount of information is always considered and that information is drawn out from the original document that is huge in size. In other words, we can say that we pluck out the summary of the main document. A summary of any document is defined as a collection of essential data by collecting the brief statements accounting the main points of the original document. Therefore, Summarization of a text is a procedure of separating or getting the relevant data out of a very large document[5]. It is the process of shortening the text document by using various technologies and methodologies to create a coherent summary including the major points of the original document. There are various methods by which the summarization process can be carried out.

While most summarization systems focus on only natural language processing (NLP), the opportunity to jointly optimize the quality of the summary with the aid of automatic speech recognition (ASR) and computer vision (CV) processing systems is widely ignored. On the other hand, given a news event (i.e., news topic), Text data are generally asynchronous in real life[7][8]. Thus, Text summarization faces a major challenge in understanding the semantics of information. In this work, we present a system that can provide users with textual summaries to help to acquire the gist of asynchronous data in a short time without reading documents from beginning to end. The purpose of this work is to unite the NLP with machine learning techniques to explore a new framework for mining the rich information contained in multi-modal data to improve the quality of Text summarization[9].

II. Review of Literature

P. Sinha, S. Mehrotra, and R. Jain, "Summarization of personal photologs using multidimensional content and context," in Proc. 1st ACM Int. Conf. Multimedia Retrieval, 2011, p. 4.

Proposed methods to compute quality, diversity and coverage properties using multidimensional content and context data. The proposed metrics which will evaluate the photo summaries based on their representation of the larger corpus and the ability to satisfy user's information needs. Advantages are: The greedy algorithm for summarization performs better than the baselines. Summaries help in effective sharing and browsing of the personal photos. Disadvantages are: Computation is expensive.

H. Lin and J. Bilmes, "Multi-document summarization via budgeted maximization of submodular functions," in Proc. Human Lang. Technol.: Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics, 2010, pp. 912–920.

In multi-document summarization, redundancy is a particularly important issue since textual units from different documents might convey the same information. A high quality (small and meaningful) summary should not only be informative about the remainder but also be compact (non-redundant). Advantages are: The best performance is achieved. Submodular summarization achieves better ROUGE-1 scores.

Disadvantages are: The proposed system very expensive to solve.

M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi, "Eddi: Interactive topic-based browsing of social status streams," in Proc. 23nd Annu. ACM Symp. User Interface Softw. Technol., 2010, pp. 303–312.
Eddi is a novel interface for browsing Twitter streams that clusters tweets by topics trending within the user's own feed. An algorithm for topic detection and a topic-oriented user interface for social information streams such as Twitter feeds. (1) benchmark TweeTopic against other topic detection approaches, and (2) compare Eddi to a typical chronological interface for consuming Twitter feeds. Advantages are: A simple, novel topic detection algorithm that uses noun-phrase detection and a search engine as an external knowledge base. Eddi is more enjoyable and more efficient to browse than the traditional chronological Twitter interface. Disadvantages are: Users had access to our clients for a limited time, making it difficult to extrapolate conclusions on how the tool might be used longitudinally. Users were viewing the history of their feed rather than tweets they had never seen before, making our task slightly less realistic.

P. Goyal, L. Behera, and T. M. Mcginnity, "A context-based word indexing model for document summarization," IEEE Transactions on Knowledge Data Engineering, vol. 25, no. 8, pp. 1693–1705, 2013. Proposes the novel idea of using the context sensitive document indexing to improve the sentence extraction-based document summarization task. In this paper, proposes a context sensitive document indexing model based on the Bernoulli model of randomness. Advantages are: The new context-based word indexing gives better performance than the baseline models. Disadvantages are: Need to calculate the lexical association over a large corpus.

D. Chakrabarti and K. Punera, "Event summarization using tweets," in Proc. 5th Int. AAAI Conf. Weblogs Social Media, 2011, pp. 66–73.
In this paper we argue that for some highly structured and recurring events, such as sports, it is better to use more sophisticated techniques to summarize the relevant tweets. The problem of summarizing event-tweets and give a solution based on learning the underlying hidden state representation of the event via Hidden Markov Models. Advantages are: The advantage of leveraging existing query matching technologies and for simple one-shot events such as earthquakes it works well. The HMM is able to learn differences in language models of sub-events completely automatically. Disadvantages are: The disadvantage that SUMMHMM has to account for tweet words that only occur in some of the events, but not in others.

Z. Li, J. Liu, J. Tang, and H. Lu, "Robust structured subspace learning for data representation," IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 10, pp. 2085–2098, 2015. In paper, proposes a singular Robust Structured Subspace Learning (RSSL) algorithm with the aid of integrating image knowledge and function gaining knowledge of into a joint studying framework. The learned subspace is accompanied as an intermediate area to reduce the semantic hollow between the low-degree seen capabilities and the high-stage semantics. Advantages are: The proposed RSSL enables to effectively research a robust based subspace from records. The proposed framework can reduce the noise-prompted uncertainty.

W. Y. Wang, Y. Mehdad, D. R. Radev, and A. Stent, "A low-rank approximation approach to learning joint embeddings of news stories and images for timeline summarization," in NAACL-HLT, 2016, pp. 58–68.
The paper proposes a singular matrix factorization technique for extractive summarization, leveraging the success of collaborative filtering. First to consider illustration learning of a joint embedding for textual content and snap shots in timeline summarization. Advantages are: It is straightforward for builders to set up the device in real-world packages. Scalable method for studying low-dimensional embedding's of information tales and snap shots. Disadvantages are: Only work on summarizing synchronous multi-modal content.

III. PROPOSED METHODOLOGY

Firstly, the file which is given as input is tokenized in order to get tokens of the terms.The stop words are removed from the text after tokenization. The words which are remained are considered as a key word.The key words are taken as an input for that we are attaching a part of tag to each key word.After completing this pre-processing step we are calculating frequency of each keyword like how frequently that key word has occurred from this maximum frequency of the keyword is taken.Now weighted frequency of the word is calculated by dividing frequency of the keywords by maximum frequency of the key words.In this step we are calculating the sum of weighted frequencies using cosine similarity.then we use LDA and Generate summary.

A. Advantages

1) It provides to automatically mine and summarize subtopics (i.e., divisions of a main topic) from large paragraph related to a given topic.
2) Document contents can facilitate subtopic discovery.
3) Well organizing the messy documents into structured subtopics.
4) Generating high quality textual summary at subtopic level.

B. Architecture

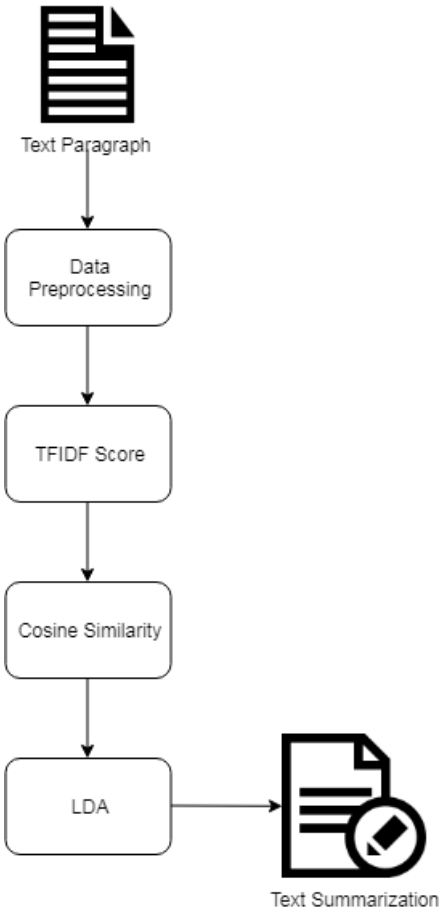

Fig. 1. Proposed System Architecture

C. Algorithms

Phase 1 – Data Preprocessing Apply preprocessing algorithms – Remove unwanted data using preprocessing algorithms.

Phase 2 – TFIDF TF: Term Frequency, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization.

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

IDF: Inverse Document Frequency, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as ”is”, ”of”, and ”that”, may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:

IDF(t) = loge(Total number of documents / Number of documents with term t in it).

Example:

Consider a document containing 100 words wherein the word cat appears 3 times. The term frequency (i.e., tf) for cat is then (3 / 100) = 0.03. Now, assume we have 10 million documents and the word cat appears in one thousand of these. Then, the inverse document frequency (i.e., idf) is calculated as log(10,000,000 / 1,000) = 4. Thus, the Tf-idf weight is the product of these quantities: 0.03 * 4 = 0.12.

Phase 3 – Cosine similarity weight Calculate cosine similarity of sentences. Remove duplicate sentences using cosine similarity weight.

Cosine similarity is a metric used to measure how similar the documents are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance (due to the size of the document), chances are they may still be oriented closer together. The smaller the angle, higher the cosine similarity.

Example:

Here are two very short texts to compare:

1. Julie loves me more than Linda loves me
2. Jane likes me more than Julie loves me

We want to know how similar these texts are, purely in terms of word counts (and ignoring word order). We begin by making a list of the words from both texts:

Julie loves Linda than more likes Jane

Now we count the number of times each of these words appears in each text:

me 2 2
Jane 0 1
Julie 1 1
Linda 1 0
likes 0 1
loves 2 1
more 1 1
than 1 1

We are not interested in the words themselves though. We are interested only in those two vertical vectors of counts. For instance, there are two instances of ’me’ in each text. We are going to decide how close these two texts are to each other by calculating one function of those two vectors, namely the cosine of the angle between them.

The two vectors are, again:

a: [2, 0, 1, 1, 0, 2, 1, 1]

b: [2, 1, 1, 0, 1, 1, 1, 1]

The cosine of the angle between them is about 0.822..

Phase 4 - Latent Dirichlet allocation(LDA) Algorithm

1. For the topic T, draw $\varphi^{TG} \sim Dir(\lambda^{TG})$ and $\varphi^{VG} \sim Dir(\lambda^{VG})$ denote the general textual distribution and visual distribution, respectively. $Dir(\cdot)$ Is the Dirichlet distribution. Then draw $\phi^{Z} \sim Dir(\beta^{Z})$, which indicates the distribution of subtopics over the microblog collection corresponding to$T$.
2. For each subtopic, draw $\varphi_k^{TS} \sim Dir(\lambda^{TS})$ and $\varphi_k^{VS} \sim Dir(\lambda^{VS})$, $k \in \{1, 2, \ldots, K\}$, correspond to the specific textual distribution and visual distribution.
3. For each microblogM_i, draw$Z_i \sim Multi(\phi^{Z})$, corresponds to the subtopic assignment for Mi. Multi $(\cdot)$ denotes the Multinomial distribution. Then draw $\phi_i^{R} \sim Dir(\beta^{R})$ indicates the general-specific textual word distribution ofM_i. Similarly, draw $\phi_i^{Q} \sim Dir(\beta^{Q})$ indicates that for visual words.
4. For each textual word position of M_i, draw a variable $R_{ij} \sim Multi(\phi_i^{R})$:
   - If R_{ij} indicates General, then draw a word$W_{ij} \sim Multi(\varphi^{TG})$.
   - If R_{ij} indicates Specific, draw a word W_{ij} from the Z_i-th specific distribution $W_{ij} \sim Multi(\varphi_{Z_i}^{TS})$
5. The generation of visual words is similarly done as in step 4.

IV. Results and Discussion

Experiments are done by a personal computer with a configuration: Intel (R) Core (TM) i3-2120 CPU @ 3.30GHz, 4GB memory, Windows 7, MySQL 5.1 backend database and jdk 1.8. The application is dynamic web application for design code in Eclipse tool and execute on Tomcat server. Some functions used in the algorithm are provided by list of jars like standford core NLP jar for keywords extraction using POS tagger method. TalkingJavaSDK jar uses for speech to text conversion and imageio jar uses for image read and write.

Some of the parameters are considered for OCR as well as ASR for text conversion methods.

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

Precision=TP/TP+FP

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class.

Recall=TP/TP+FN

F-measure - F-measure is the weighted average of Precision and Recall.

F-measure = 2*(Precision*Recall) / (Precision+Recall)

Accuracy - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

Accuracy = TP+TN/TP+FP+FN+TN

V. Conclusion

Automatic text summarization is a complex task which contains many sub-tasks in it. Every subtask has an ability to get good quality summaries. The important part in extractive text summarization is identifying necessary paragraphs from the given document. In this work we proposed extractive based text summarization by using statistical novel approach based on the sentences ranking the sentences are selected by the summarizer. The sentences which are extracted are produced as a summarized text and it is converted into audio form. The proposed model improves the accuracy when compared traditional approach.

References

[1] Cheng, Jianpeng, and Mirella Lapata. "Neural summarization by extracting sentences and words."arXiv preprint arXiv:1603.07252 (2016).

[2] Nallapati, Ramesh, et al. "Abstractive text summarization using sequenceto- sequence rnns and beyond." arXiv preprint arXiv:1602.06023 (2016).

[3] Freitas, N., and A. Kaestner. "Automatic text summarization using a machine learning approach." Brazilian Symposium on Artificial Intelligence (SBIA), Brazil. 2005.

[4] Ferreira, Rafael, et al. "Assessing sentence scoring techniques for extractive text summarization." Expert systems with applications 40.14 (2013): 5755-5764.

[5] Gaikwad, Deepali K., and C. Namrata Mahender. "A Review Paper on Text Summarization." International Journal of Advanced Research in Computer and Communication Engineering 5.3 (2016).

[6] Fachrurrozi, M., Novi Yusliani, and Rizky Utami Yoanita. "Frequent Term based Text Summarization for Bahasa Indonesia." (2013): 30-32.

[7] Radev, Dragomir R., et al. "Centroid-based summarization of multiple documents." Information Processing Management40.6 (2004): 919-938.

[8] P. Sinha, S. Mehrotra, and R. Jain, "Summarization of personal photologs using multidimensional content and context," in Proc. 1st ACM Int. Conf. Multimedia Retrieval, 2011, p. 4.

[9] H. Lin and J. Bilmes, "Multi-document summarization via budgeted maximization of submodular functions," in Proc. Human Lang. Technol.: Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics, 2010, pp. 912–920.

[10] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi, "Eddi: Interactive topic-based browsing of social status streams," in Proc. 23nd Annu. ACM Symp. User Interface Softw. Technol., 2010, pp. 303–312.

[11] P. Goyal, L. Behera, and T. M. Mcginnity, "A context-based word indexing model for document summarization," IEEE Transactions on Knowledge Data Engineering, vol. 25, no. 8, pp. 1693–1705, 2013.

[12] D. Chakrabarti and K. Punera, "Event summarization using tweets," in Proc. 5th Int. AAAI Conf. Weblogs Social Media, 2011, pp. 66–73.