ISSN: 2586-761x (Print)
ISSN: 2635-5817 (Online)
Vol. 02, No. 01, June 2019
International Journal of Advanced Social Sciences
Source: http://ictaes.org
Manuscript received: January 20, 2019 ; Revised: March 10, 2019 ; Accepted: April 20, 2019

Salient Sentence Extraction of Nepali Online Health News Texts

Robin Ranabhat, Amit Upreti, Bidhan Sangpang, and Shoaib Manandhar

Department of Computer Science and Engineering, Kathmandu University, Dhulikhel, Kavre, Nepal

robinnarsingha123@gmail.com, a.u.aua937@gmail.com, bidhan.rai@student.ku.edu.np, shoaib.manandhar@student.ku.edu.np

Abstract

Exponential growth of the Nepali language content has made information processing and retrieval prominent. We cannot possibly create summaries of all of the text manually; there is a great need for automatic methods. Textual information in the form of digital documents quickly accumulates to huge amounts of data. Most of this large volume of documents is unstructured: it is unrestricted and has not been organized into traditional databases. Processing documents is therefore a perfunctory task, mostly due to the lack of standards. We present you with a prototype android application to aid users in reading news by providing a sentence-based summary of the Nepalese news.

Keywords: *extractive summarization, topic modeling, TextRank (graph algorithm)*

1. INTRODUCTION

The dramatic growth of information and documents in the Internet has demanded exhaustive research in the field of automatic text summarization. Automatic text summarization is the task of producing a concise and fluent summary while preserving key information content and overall meaning [1]. Examples include search engines generate snippets preview of documents.

In this paper, a Graph based sentence scoring technique is accessed, Text rank to perform extractive summarization, where, we produce summaries by choosing a subset of the sentences in the original text. These summaries contain the most important sentences of the input [2].

The extractive text summarization algorithm was implemented in an android application to aid reading Nepalese news articles. We hope to research how better can extractive summarization be used in case of natural Nepalese text and implement the results through an android application.

In recent years, numerous approaches have been developed for automatic text summarization and applied widely in various domains. The trend of reducing the information in this hectic time where people always seem to be out of time.

Corresponding author : Robin Ranabhat
Author's affiliation : *Kathmandu University,Nepal*
Email: *robinnarsingha123@gmail.com*

2. Related Works

Sentence extraction is an important first step for text summarization that shows how sentence-based extraction works. And a large method body of algorithms are present for sentence extraction. Previous approaches include supervised learning, vectoral similarity computed between an initial abstract and sentences in the given document, or intra-document similarities [3]. It is also notable the study reported in [4] discussing the usefulness and limitations of automatic sentence extraction for summarization, which emphasizes the need of accurate tools for sentence extraction, as an integral part of automatic summarization systems. Table 1 shows an evaluation of TextRank summary.

Table 1. Evaluation of TextRank Summary

System	ROUGE score – Ngram(1,1)		
	basic (a)	stemmed (b)	stemmed no-stopwords (c)
S27	0.4814	0.5011	0.4405
S31	0.4715	0.4914	0.4160
TextRank	**0.4708**	**0.4904**	**0.4229**
S28	0.4703	0.4890	0.4346
S21	0.4683	0.4869	0.4222
Baseline	*0.4599*	*0.4779*	*0.4162*
S29	0.4502	0.4681	0.4019

3. Dataset and Pre-processing

We created the ’25nepaliarticles’ corpus from scratch. The news articles on different categories published on nepalihealth.com were parsed and scrapped. The corpus includes 25,000 articles on 9 different categories. Fig. 1 shows the distribution of nepali health article for each category. Most of the data was from Sports News category while the least data was from technology category. Pre-Processing the news article data involved two major steps. First, the news article collected were cleaned by stripping all unwanted tags, symbols from the document like ‘, . ।’ etc. and other text such as ‘author name, published date etc. Also, Redundant word removal like ‘छ’, ‘पनि’, ‘भएको’, ‘गरेको’ from document. Then the news article was organized in a csv file with labels title article and category.

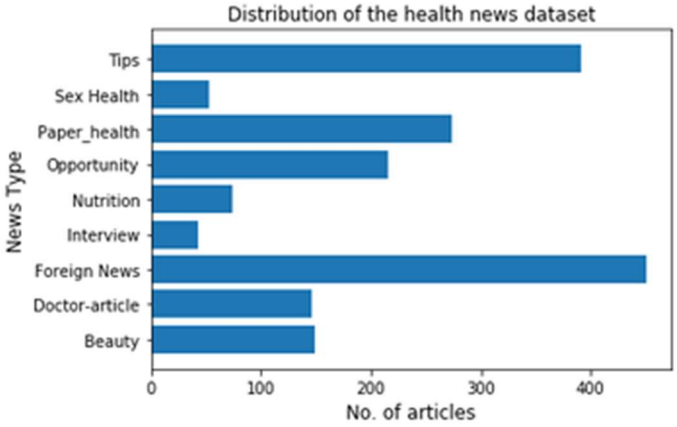

Figure 1. Distribution of the 25 Nepali article Dataset

4. METHODOLOGY

Our extractive summarization system identifies the most important sentences in the input, which is a single news article and collectively string them together to form a summary.

We distinguish three prime tasks performed by our system:

- Preprocessing the raw nepalese text
- Rank the sentences
- Selecting a summary consisting of several sentences in order of their rankings.

4.1 Preprocessing raw text

Preprocessing the raw data involved two major steps:

- Stripping all unwanted tags, symbols from the document
- Redundant word removal like 'छ', 'पनि', 'भएको', 'गरेको' from document.

4.2 Extract important sentences with TextRank

Eventually, the summarizer system selects the top k most important sentences to produce a summa ry. The TextRank algorithm enables application of graph-based sentence ranking to natural language te xt by marking each sentence as vertices of the graph and maintaining similarity score [5].

We used TextRank to rank the sentences from the given news article. Then, the top three sent ences were marked as summary for a given news. Some of the best results of our summarization app roach are highlighted below in Table II.

4.3 Summary Evaluation with Topic Modeling

Evaluation of a summary is a difficult task because there is no ideal summary for a document, or a collection of documents and the definition of a good summary is an open question to large extent [6]. To be able to to automatic summary evaluation, we must conquer the following difficulties:

- It is fundamental to decide and specify the most important parts of the original text to preserve.
- These evaluators must automatically identify these pieces of important information in the candidate sum mary, since this information can be represented using disparate expressions.
- The readability of the summary in terms of grammaticality and coherence must be evaluated.

Since human evaluation is not practical and we lacked the test set (already summarized news a rticles) to use widely celebrated methods like Recall-Oriented Understudy for Gisting Evaluation (ROU GE), we had to decide our own set of metrics. To cope with the first difficulty, we use an unsupervi sed machine learning algorithm, latent dirichlet allocation (LDA) to find important words from each c ategory.

For each category of news, we need a metric to find how good our summary is as we lacked the human evaluated gold standard. To overcome that, for each category of news, we defined our ow n metric where we calculated the similarity between the words generated in the summary and the im portant words of that category.

For example, if it's a sport news, we evaluated the similarity between the summarized news an d important words of sport category. An unsupervised machine learning algorithm called latent dirichle t allocation (LDA) was used to find important words from each category. Latent Dirichlet allocation

[7] represents documents as random mixtures over latent topics, where each topic is characterized by a distribution over words. Each word wt in a corpus w is assumed to have been generated by a late nt topic zt, drawn from a document-specific distribution over T topics. We ran the LDA algorithm ov er previously scraped Nepalese news article data to generate list of important words for each news ca tegory as shown below:

Table II. Average summary score of each category

News Category	Summary Match Score (%)
Nutrition	3.38
Paper Health	2.66
Beauty	4.802
Tip	3.811
Sex Health	2.4439
Foreign Health	3.106
Interview	2.64
Doctor Article	2.37
Opportunity	5.5865

5. RESEARCH AND DISCUSSION

Our approach to use the graph-based summarization presented better than expected results. The news of category 'Tips' and 'Nutrition' provided decent results. The information provided in these categories were based in fact and specific. So, the algorithm was successful in extracting the information of value.

We found that, for the news category 'interview', it's quite hard to make sense just by reading the most important sentences only. These categories contained information on wide subject range, opinions of a group because of which some of the most important sentence couldn't address the overall gist of whole news article, unlike in 'Nu' news.

We also noticed that, most of the times, the most important sentence included the first sentence of the news article. This suggested that, the first line of a news article is very significant in carrying the overall gist of the article.

For the evaluation of the summary, we couldn't apply any present summarization metric because we didn't have the test data against which we could measure the accuracy of our summary. So, we came up with our own metric. After finding the important words for each category as discussed above, we scored each summary as

$$score = \frac{N}{M}$$

where, N is the number of matched words between summarized news and our LDA generated words from Table III and M is the total number of words in the summarized new article.

Since no such metric has been established before, we can't know how much score can be regarded as a standard score for a decent summary. But what we can infer is, summaries with low score value are most of the time worse than summaries with high score values when they are human evaluated.

Generally, with score value of 10 or higher, we got summaries that were quite good but the main problem we got was that, for most of the case, the summaries generated had very low score values. Also, although not a drawback, for very small news, algorithm couldn't find any relevant sentences. But since news were smaller from the start, we concluded there was no point in trying to find important sentence anyways.

Table III. (a) High Summary score

Article	Summary	Category	Summary Score
कुनै पनि स्तनधारी कुनैपनि स्तनधारी प्राणीका लागि दुध अतिनै आवश्यक पदार्थ हो । अझ मानव जीवनका लागि त झनै महत्वपूर्ण पदार्थ हो । कुनै पनि शिशुका लागि जन्मेको ६ महिना सम्म उसको सम्पूर्ण खानेकुरा भनेकै दुध मात्र हो ...त्यसैले दुधमा हुने विभिनन पदार्थलाई आवश्यकता अनुसार विभिन्न परिकारमा परिवर्तन गरी फाईदा लिन सकिन्छ ।,(डा. ज्ञवाली त्रिभुवन विश्वविद्यालय चिकित्साशास्त्र अध्ययन संस्थान अन्तर्गत महाराजगञ्ज क्याम्पसमा कार्यरत छन)	शरीरका लागि आवश्यक पदार्थ दुधबाटै पाइने हुनाले नियमित रुपमा दुध पिउने व्यक्ति शारीरिकरुपमा तन्दुरुस्थ र निरोगी हुन्छ । दुधमा पाइने प्रोटिनले शरीरको वृत्तिविकासको लागि काम गर्छ। तुलनात्मक रुपमा गाईको भन्दा भैँसीको दुधमा कार्वोहाइट्रेड र बोसोको मात्रा बढी हुन्छ । त्यसैले प्याकेट दुधको खरीदमा त्यसको प्याकेटमा लेखिएको तत्वहरुको समेत अध्ययन गर्नु जरुरी हुन्छ ।	Nutrition	5 % Match Score
छ छ । तर । , त आठ तर । ... थपघट ।	खाना खाएपछि एक गिलास पानी पिएको राम्रो यसले शरीरको पाँचन शक्तिलाई थप सहयोग पुरयाउँछ । तर विज्ञहरुले यो पनि भनेका छन् कि शरीर, बाताबरण, काम अनुसार पनि पानी पिउने मात्रा र समय फेरबदल गर्न सकिन्छ ।	Tips	4.125 % Match Score

Table III. (b) Low Summary score

Article	Summary	Category	Summary Score
र १० । । . १०-१५ । ।	१०-१५ र	Beauty	1.3 %
‘ ’ । लगभग ४६ । ४६ छ । : , . समय खपतदर कम छ ।	मैले माथि नै भने हामी आधा भन्दा अलि कम हिस्सा ओगटेका छौँ बाहिरबाट आउनेले आधाभन्दा केही बढी हिस्सा ओगटेका छन् । १० बर्ष पछिको भिजन हेर्ने हो भने नेपालमा अहिले ५५वटा औषधि उद्योग छन् केही नभएपनि त्यो संख्या बढेर ७० पुग्नेछ	Interview	3.18 %

5. CONCLUSION

The increasing growth of the Internet has made a huge amount of information (health news) available. It will be very helpful for general public a decent information retrieval system. Thus, there is an immense need for automating information extraction in this age of information overload.

In this paper, we emphasized an extractive approach for important information retrieval from health articles. We described some of the most extensively used methods such as TextRank for sentence scoring and topic modeling for summary evaluation purposes. Although we couldn't use deep learning to train a summarization algorithm that can generate semantic summary based due to lack of training data, we hope to

train a title generation model for Nepalese news article using neural networks later on for which training data can be collected.

ACKNOWLEDGEMENT

We would like to acknowledge Assoc. Prof. Dr. Bal Krishna Bal, Lead Researcher at Information and Language Processing Research Lab, Department of Computer Science and Engineering, Kathmandu University and Assist. Prof. Mr. Manoj Shakya, Lead Researcher at Digital Research and Learning Lab, Department of Computer Science and Engineering, Kathmandu University and Assist. Prof. Mr. Sushil Shrestha, Lead Researcher at Digital Research and Learning Lab, Department of Computer Science and Engineering, Kathmandu University, for their valuable input during research and development of this research work.

REFERENCE

[1] C. Lin and E. Hovy, "Automatic evaluation of summaries using N-gram co-occurrence statistics", Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03, 2003.

[2] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D., J. B. and K. Kochut. (2017), Text Summarization Techniques: A Brief Survey, International Journal of Advanced Computer Science and Applications, Vol. 8, No. 10.

[3] Salton, G., Singhal, A., Mitra, M. and Buckley, C. (1997). Automatic text structuring and summarization. Information Processing & Management, 33(2), pp.193-207.

[4] Lin, C. and Hovy, E. (2003). Automatic evaluation of summaries using N-gram co-occurrence statistics. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03.

[5] R. Mihalcea and P. Tarau, (2004). TextRank: Bringing Order into Texts

[6] C. Lin and E. Hovy. (2003). Automatic evaluation of summaries using N-gram co-occurrence statistics", Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03, 2003.

[7] H. Wallach, (2006). Topic modeling", Proceedings of the 23rd international conference on Machine learning - ICML 2006.