Nepali News Headline Generation Using Extractive Method

Ashutosh Chapagain, Anup Sedhain and Krishu Thapa

Abstract

Headline provides the gist of the news which help the reader to get a better intuition of the whole news. This paper presents a new alternative to existing Headline Generation Techniques. Although, the paper is demonstrated on Devanagari script (Nepali language), it can be easily adopted for other languages. The main aim is to construct headline from key terms for saving the interpretation and reading time of reader.

1 Introduction

Headline generation is within the category of the text summarization. Text summarization is the process of distilling the most important information from a text to produce an abridged version for a particular task and user (definition adapted from Mani and Maybury (1999)). One crucial architectural dimension for text summarizers is whether they are producing an abstract or an extract. The simplest kind of summary, an extract, is formed by selecting (extracting) phrases or sentences from the document to be summarized and pasting them together. By contrast, an abstract uses different words to describe the contents of the document. [1]. A special application of text summarization is generating very short summarizes from input text, or headlines from news articles and documents(thesis), and is the focus of this work. In this paper, we would like to modify the current way (extractive) of headline generation and automatically generate headlines from the text of news articles.

This work deals with Nepali news data scraped from onlinekhabar.com. To limit the vocabulary size, only technology news are taken into consideration. [3] is closely related to this where the major focus was to frame headline from key terms and candidate phrases using NLP techniques. Our work focuses equally on summarization techniques as well to frame headline of a given news. [5] are the abstract methods which use simple attention based models and Recurrent Neural Networks respectively. The training data consisted of 5.5M news articles with 236M words. Due to the lack of hand-annotated dataset available, extractive methods were used to frame headline of Nepali News Text. Also, [1] recommended to use extractive task over abstractive one.

Sentences are scored using the TextRank algorithm and top 3 sentences are chosen. Independently, Rapid automatic keyword extraction technique is used to extract the keywords of the news.[2] Sentences which contain the most keywords is selected for further processing and is compressed to a headline. It is clear that a program can rank one sentence higher than the other simply by comparing n-gram matches between each keywords and summarized sentence[6]. We will evaluate generated news headlines with ROGUE [6]. In general, ROGUE measures how much the words in the machine-generated headlines appeared in the human reference headlines.

The remaining paper is organized as follows: section 2 contains methodology while the next section contains the result where we evaluate it with a ROGUE measure which is an evaluation metrics for evaluating summaries[1].

2 Methodology

The news headline formation includes the following major steps:

- Top-3 Sentence Extraction
- Key term/keyphrase Extraction

Figure 1 shows the entire pipeline used for Nepali News Headline Generation Using Extractive Method.

2.1 Preprocessing

The news articles scraped from onlinekhabar.com, needed to be preprocessed before it could be used.The processes involved in the preprocessing were stopwords removal, special characters removal,stemming, sentence tokenization.

2.1.1 Stemming

Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma.The goal of stemming is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.Nepali, being a highly inflectional and derivational language, a single word can represent various grammatical forms and meanings. For example a verb root लेख्(lekh) can show different forms such as: लेख्छु(lekh-chu),लेख्छस्(lekh-chas),लेखछेस्(lekh-ches), लेख्छ(lekh-cha), लेखी(lekh-i), लेख्यो(lekh-yo), लेखे(lekh-e).So it is necessary to perform stemming in order to obtain the desired result with good accuracy.

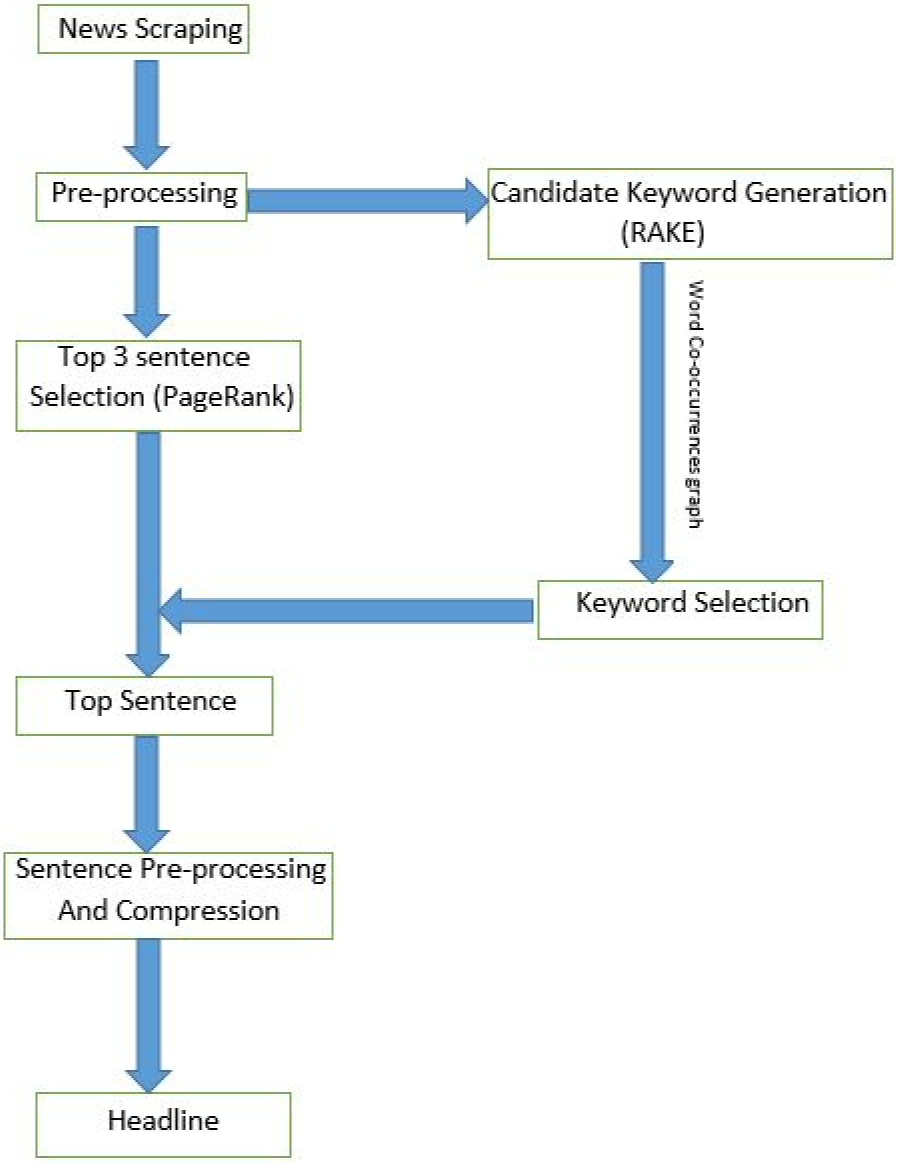
News Scraping
Pre-processing
Candidate Keyword Generation (RAKE)
Top 3 sentence Selection (PageRank)
Word Co-occurrencesgraph
Keyword Selection
Top Sentence
Sentence Pre-processing And Compression
Headline

Figure 1 : Steps for Headline generation using extractive methods

2.1.2 Special Characters Removal

An article,along with texts contains different non-textual special characters like '[?!/(){}\[\]\|@,;\'\']'. They don't provide any value to the outcome of the analysis. Removing these characters is a part of text cleaning which enables the transformation of unstructured data to an understandable form.

2.1.3 Stopwords Removal

Stop words are the words used in defining the structure of sentences. These are the most frequent words in a corpus but they do not provide any value to the outcome of an analysis. They only take up extra processing time. Removing stop words during pre-processing lets us to focus more on the important words that provide value to the outcome of the analysis.Stop words can include language specific determiners, conjunctions and postpositions. Also, it can include other common words like names, places and temporal words.Some nepali stopwords are: determiners (यो(yo), यी(yi)), conjunctions(र(ra), तथा(tatha)),postpositions(ले(le),बाट(bat) लाई(lai:)), etc.

2.1.4 Sentence Tokenization

Tokenization the process of splitting the given string into units called tokens. A token is a sequence of character, usually word or sentence that is semantically significant for text analysis.In cases where we want our tokens to be sentences, our possible token boundary is either पूर्णविराम(।) or प्रश्न चिन्ह(?) or विस्मयादिबोधक(!). So, by splitting the given text at पूर्णविराम(।) or प्रश्न चिन्ह(?) or विस्मयादिबोधक(!), sentence-level tokenization can be achieved.

2.2 Keyword Extraction using RAKE

Rapid Automatic Keyword Extraction (**RAKE**) is an unsupervised, domain-independent, and language-independent method for extracting keywords from individual documents. The key terms and candidate phrases are extracted from input news article and it is an extractive method for single document keyword generation. The input parameters for RAKE comprise a list of stop words (or stoplist), a set of phrase delimiters, and a set of word delimiters. RAKE uses stop words and phrase delimiters to partition the document text into candidate keywords. Co-occurrences of words within these candidate keywords allow us to construct a word co-occurrence graph. Word associations are thus measured in a manner that automatically adapts to the style and content of the text, enabling adaptive and fine-grained measurement of word co-occurrences that will be used to score candidate keywords.

2.2.1 Scoring Technique

After every candidate keyword is identified and the graph of word co-occurrences is complete, a score is calculated for each candidate key-word and defined as the sum of its member word scores. We evaluated several metrics for calculating word scores, based on the degree and frequency of word vertices in the graph: word frequency (freq(w)), word degree (deg(w)), and ratio of degree to frequency (deg(w)/freq(w)).

Each of these provide a solid ground for providing scores to the keywords. The word degree i.e. deg(w) favors words that occur often and in longer candidate keywords. Words that occur frequently regardless of the number of words with which they co-occur are favored by freq(w). Finally, words that predominantly occur in longer candidate keywords are favored by deg(w)/freq(w), the ratio between the two. More about this technique could be found here[2].

2.3 Top three Sentence Extraction

A graph-based ranking algorithm (TextRank) is applied to rank words based on their associations in the graph, and then top ranking words are selected as keywords. TextRank is an extractive and unsupervised text summarization technique. The process for selecting top-3 algorithm is shown in figure 2 . More about TextRank algorithm can be found at [7].

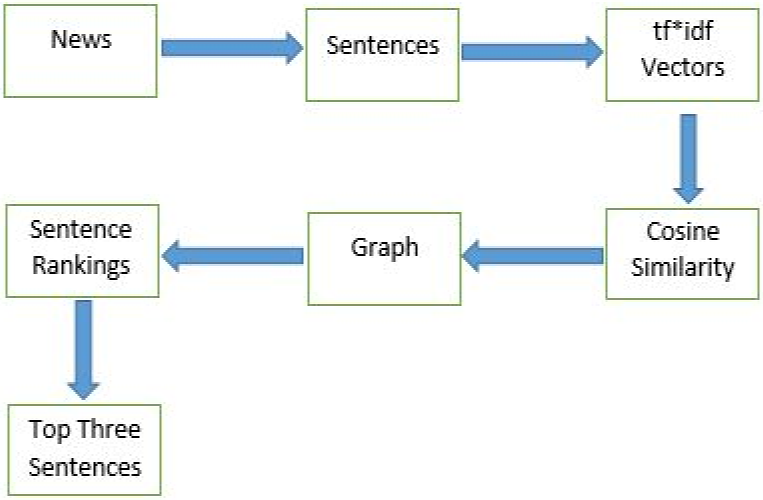

Figure 2 : Steps for top 3 sentence extraction using Textrank Algorithm

2.3.1 Similarity Between Top Sentences and Keywords

Edit distance (Levenshtein distance) is calculated among the top three sentence and the top scored keywords from RAKE. The sentence which had the most similarity to the top-scored list of keywords are the most informative sentences.

2.4 Finding accuracy using Rogue

The primary programming task for a ROGUE implementer is to compare n-grams of the candidate with the n-grams of the reference headline and count the number of matches[6]. These matches are position-independent.The more matches,the better the candidate translation is. Both ROGUE-I(recall) and ROGUE-II(precision) matches are calculated.

3 Result and Evaluation

3.1 Result comparison

Generated Heading	ctual Heading	OGUE Precision	OGUE Recall
1. कम्पनी प्रमुख वित्तिय अधिकृत लू वार्षिक तलब ८१ प्रतिशत वृद्धि उन गत वर्ष अढाइ अर्ब रुपैयाँ बढि तलब भत्ता बुझे	1.एप्पल सिइओ वर्षदिन तलब भत्ता १ अर्ब , सबैभन्दा तलब खाने महिला हाकिम	0.1578	0.25
2. इक्रोब्लगिंग साइट ट्विटर ट्विट्स रहँदै १४० क्यारेक्टर्स सी बढाएर छिटै १० पूराउने तयार गरिरहे	2.ट्विटर बढाउँदैछ अक्षर सी , छिटै १० अक्षर ट्विट सकिने	0.285	0.444
3. टु फ्याक्टर अथेन्टिफिकेसन परेर तपाइँ अनलाइन अउण्ट लगइन पासवर्ड कमजोर राख्नुभयो प्रत्युत्पादक बन्नेछ	3.पासवर्ड सबैभन्दा सुरक्षित ?	0.076	0.25
4.अमेरिकी इन्टरनेट भिडियो सर्भिसनेटफ्लिक्स सेवा	4.नेटफ्लिक्स नेपालसहित बिश्वभर हेर्न सकिने	0.3	0.4

विश्वब्यापी बनाए			

3.2 Difficulties and Shortcomings

- Anaphora resolution was the biggest difficulty in our model. The pronoun (कम्पनी) appeared instead of the proper-noun(एप्पल).
- Words with higher lexical similarity were not recognized as same word(words like बिश्वभर and विश्वब्यापी).
- As the generated sentence was the part was the part of actual news, the sentences were very long and it lead to low recall.

3.3 Future Work

- Implementation of all stemming rules to increase the ROGUE measure of our model.
- Implementation of Named Entity Recognition for anaphora resolution.
- Sentence compression algorithm to be used to decrease the size of generated headline without losing its meaning.

4. Conclusion

In this particular implementation of single document summarization, we used two extraction algorithms for getting better results. We simply extracted words from one and sentences from the other that contained most number of keywords hence giving it the maximum score. This made the procedure more inclined towards producing sentences with greater content and at the end making the single sentence rich for it being used as a headline for any news coverage.

References

[1]D. Jurafsky and J. Martin, *Speech and language processing*. Upper Saddle River, N.J.: Pearson Prentice Hall, 2009.

[2]S. Rose, D. Engel, N. Cramer and W. Cowley, "Automatic Keyword Extraction from Individual Documents", *Text Mining*, pp. 1-20, 2010. Available: 10.1002/9780470689646.ch1 [Accessed 16 June 2019].

[3]U. Shrawankar and K. Wankhede, "News Headline Building using Hybrid Headline Generation Technique for Quick Gist", *International Journal of Natural Computing Research*, vol. 6, no. 1, pp. 36-52, 2017. Available: 10.4018/ijncr.2017010103.

[4]P. McBurney, C. Liu and C. McMillan, "Automated feature discovery via sentence selection and source code summarization", *Journal of Software: Evolution and Process*, vol. 28, no. 2, pp. 120-145, 2016. Available: 10.1002/smr.1768.

[5] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks.CoRR, abs/1409.3215, 2014.

[6] Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. BLEU: a method for automatic evaluation of machine translation. ACL-2002: 40th Annual Meeting of the Association for Computational Linguistics. Pp. 311-318.

[7]A. implementation), "Introduction to Text Summarization using the TextRank Algorithm", *Analytics Vidhya*, 2019. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/. [Accessed: 16- Jun- 2019].