Framing News Headline from Key Terms Using NLP

Kranti Wankhede* and Urmila Shrawankar**
*Student Researcher
w.kranti666@gmail.com
**IEEE Member
urmila@ieee.org

Abstract: News headline provides the gist of news article which helps reader to understand whole idea of news without reading it. In this paper, news headline is formed by using key terms and candidate phrases which are retrieved from input news article along with Natural Language Processing (NLP) techniques. Candidate phrases are extracted from input news article by using Keyphrase Extraction Algorithm (KEA). The word dictionary for various kinds of news articles along with some more techniques of keyword extraction are used as criteria for selecting keywords. Parsing technique and sentence compression algorithm are used for construction of proper news headline from leading sentences. Headline gives the brief idea of lengthy news article. The main aim is to construct headline from key terms for saving the interpretation and reading time of reader.

Keywords: Sentence compression algorithm, headline generation; Natural Language Processing (NLP); Key term extraction; candidate phrase extraction.

Introduction

Nowadays, due to busy schedule, readers are unable to read lengthy news articles. Therefore, headlines are required for such news articles for proper and quick understanding of main content of news article. The lengthy news stories contain more information than average required length. Usually, headlines can be constructed by analyzing and understanding the main concepts of news text. The headline helps to save reading as well as interpretation time of entire news article. The headline formation can be done by using key terms/keyphrase extraction and various techniques of Natural Language Processing (NLP).

The key terms and candidate phrases are extracted from input news article by using various techniques of keyword extraction [1]. The headline generation technique is also considered as a tool for extraction of vital points of text as well as information overload minimization [2]. The abstractive and extractive techniques [3, 4, 5] are used conventionally for headline generation. The headline, which is framed by using key terms, is very effective because key terms contain vital information of the piece of writing.

The research work given in this paper is extension of the work in [6]. The objective of this work is to frame headline using NLP techniques and key terms/ keyphrases extracted from complete news article. This work deals with only English news stories. The standard dataset of various kinds of BBC news such as politics, sports etc. is used in this work. Any online news article can also be taken as input for the system. The pre-processing is done on the input news article which includes sentence segmentation, tokenization, stemming etc. The headline is formed from leading sentences of news story by applying parse tree generation technique of core NLP [7] and sentence compression algorithm. The word dictionary and some more approaches are used for keyword extraction [1]. The candidate phrases are retrieved with the help of Keyphrase Extraction Algorithm (KEA) [8, 9, 10, 11] from input news story. The leading sentences compression is done by using sentence compression algorithm [7] which results into compressed sentence generation. Then by further processing headline can be formed from compressed sentences.

The remaining paper is arranged as below. The proposed work is discussed in the following section followed by the conclusion of the system.

Proposed Work

Procedure of project

The news headline formation includes the following major steps:

- Key term/keyphrase Extraction
- Parsing of leading sentences
- Generation of compressed sentences

ISBN-13:978-93-5260-126-4, ISBN-10:93-5260-126-2

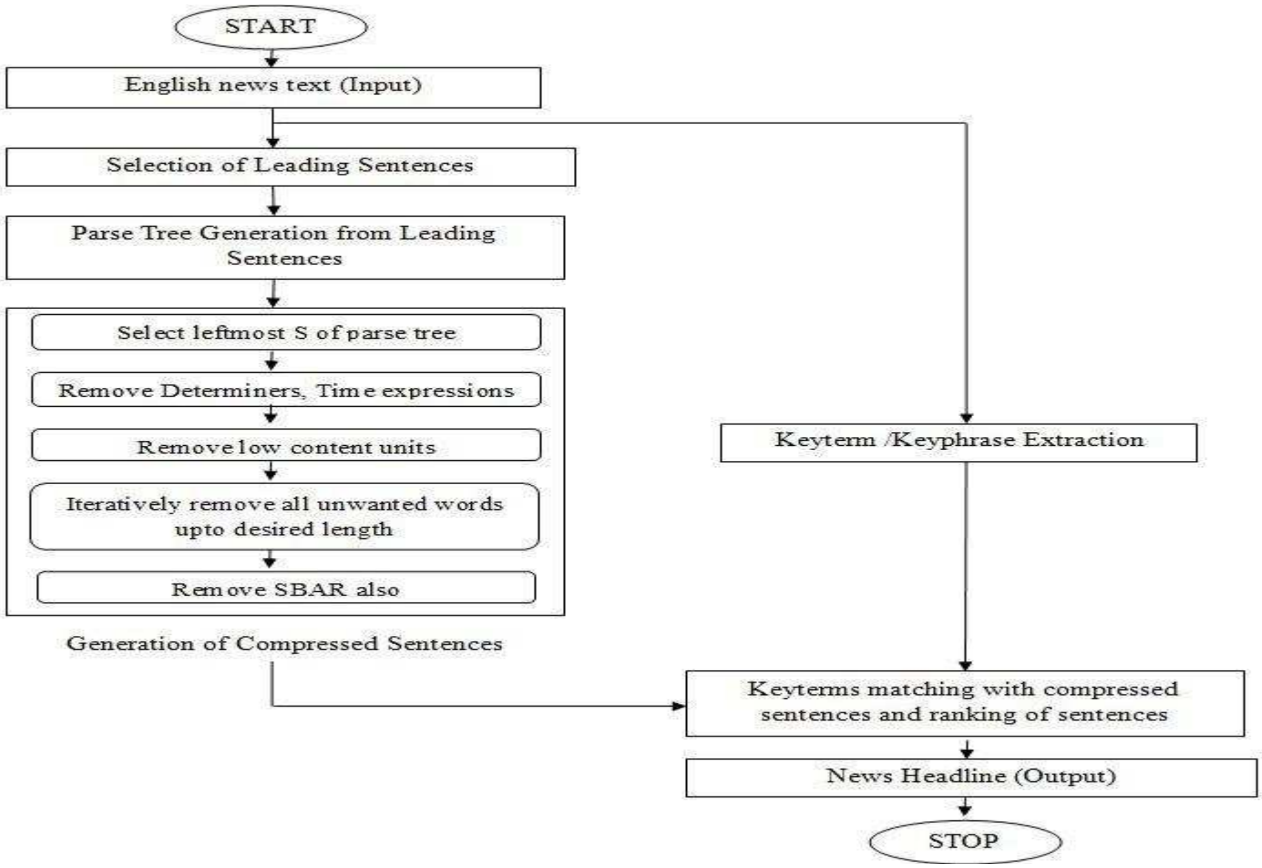
START
English news text (Input)
Selection of Leading Sentences
Parse Tree Generation from Leading Sentences
Select leftmost S of parse tree
Remove Determiners, Time expressions
Remove low content units
Iteratively remove all unwanted words upto desired length
Remove SBAR also
Generation of Compressed Sentences
Keyterm /Keyphrase Extraction
Keyterms matching with compressed sentences and ranking of sentences
News Headline (Output)
STOP

Figure1. Major steps for headline generation from key terms using NLP techniques

Keyterm/keyphrase Extraction

The news headline can be framed by using key terms and candidate phrases from input news paragraph. A plenty of techniques are available for key terms and candidate phrase extraction [12, 13] which includes machine learning approaches, statistical methods [14], unsupervised and supervised approaches [15]. In this system the key terms are retrieved by using various features like Term Frequency, Numerical Data, Term Count, Temporal Data etc. The word dictionary for different kinds of news is also used for key terms extraction. The dictionary words, if found in the input news, are picked out as key term from input news.

Many of the news headlines contain candidate phrases which gives vital information of news event. For candidate phrase extraction various methods and techniques are available. Keyphrase Extraction Algorithm (KEA) [8, 9] is one of these available techniques, used in this work for retrieving keyphrases.

The steps of KEA for keyphrase selection are as follows:

- The tokens are separated by tokenization in preprocessing steps.
- The punctuation marks, numbers, brackets are removed.
- The next step is candidate phrase identification, in which KEA uses overall subsequences for consideration of keyphrase in each and every line.
- Then it recognizes the candidate phrases of suitable length, not more than three words and it should not start or end with stop word.
- Then it extracts keyphrases according to defined grammar patterns.

ISBN-13:978-93-5260-126-4, ISBN-10:93-5260-126-2

Parsing of leading sentences

According to [16], the leading paragraph of news story is most informative and maximum words of the headline are present in first sentence of news story. The compression of first leading sentence results into headline generation for news articles.

Using this concept, the proposed system generates parse tree of the leading sentences of news article. In the parsing [17] phase, the tags are given to each word of the sentence and parse tree is generated from input lead sentences.

For example: Stanford parser is an efficient parser for generating parse tree.

Parse Tree:

```
(ROOT
  (S
    (NP (JJ Stanford) (NN parser))
    (VP (VBZ is)
      (NP
        (NP (DT an) (JJ efficient) (NN parser))
        (PP (IN for)
          (NP (VBG generating) (NN parse) (NN tree)))))
           )))
```

Generation of compressed sentences

The leading sentence compression is performed by using compression algorithm [7, 18]. The compressed sentences are formed from generated parsed sentences by using sentence compression algorithm. The compression algorithm begins by removing unwanted words of sentence such as determiners, time expressions and other some low content words. Many more compression rules are then applied for eliminating larger constituents of the parse tree up to desired headline length is achieved. This is important step for the headline generation [19, 20]. The steps involved in generating compressed sentence from parsed sentence are given as below:

- Select the leftmost S of generated parse tree.
- Remove all determiners and time expressions.
- Remove all low content units such as quantifiers (e.g. many, each, some), possessive pronouns (e.g. hers, their, ours) and deictic (e.g. this, these, those).
- Then iteratively remove all constituents until the desired length is reached.
- Remove the trailing SBAR also.

The following example helps to illustrate the sentence compression algorithm.

Lead Sentence

The dollar has hit its highest level against the euro in almost three months after the Federal Reserve head said the US trade deficit is set to stabilize.

Parsed sentence

(ROOT (S (NP (DT The) (NN dollar))(VP (VBZ has) (VP (VBN hit) (NP (NP (PRP$ its) (JJS highest) (NN level)) (PP (IN against) (NP (NP (DT the) (NN euro)) (PP (IN in) (NP (RB almost) (CD three) (NNS months))))))(SBAR (IN after)(S (NP (DT the) (NNP Federal) (NNP Reserve) (NN head))(VP (VBD said) (SBAR (S(NP (DT the) (NNP US) (NN trade) (NN deficit)) (VP (VBZ is)(VP (VBN set) (PP (TO to)(NP (NN stabilize.))))))))))))))

Compressed sentence after removal of tags

dollar has hit highest level against euro in three months.

In the above example, a lead sentence is taken and parse tree is generated for that leading sentence. Then by applying all the steps of sentence compression algorithm the above compressed sentence is formed from the parsed sentence. In this manner, all leading sentences of leading paragraph are parsed and compressed sentences are generated from them. Then the key terms which are retrieved from input news article are matched with these generated compressed sentences. The sentence having highest number of key terms are taken out as headline and further post-processing is done on that headline. The post-processing includes syntax correction, tense identification, grammar checking etc.

Conclusion

This research work frames headline from key terms and candidate phrases using NLP techniques. The key terms are retrieved using word dictionary and some more features such as temporal and numerical data, Term Frequency, Term Count etc. The candidate phrases are extracted by applying KEA algorithm which further used in the headline formation. The NLP's parsing technique is used for parse tree generation of leading sentences. The compressed sentences are constructed from leading sentences of news stories by using sentence compression algorithm. The extracted key terms are then matched with compressed sentences and ranking is performed on compressed sentences. The compressed sentence having more number of keywords is taken out as most possible headline of the input news story. Then this headline undergoes some post-processing steps like grammar checking, tense identification, syntax correction etc. which results into proper news headline.

References

[1] M. Habibi , and A. Popescu-Belis ,"Keyword Extraction and Clustering for Document Recommendation in Conversations", IEEE Transactions On Audio, Speech, And Language Processing, Vol. 23, No. 4, 2015, pp.746-759.

[2] K. Kaikhah, "Automatic text summarization with neural networks", Second International IEEE Conference on Intelligent System, 2004, pp. 40 – 45.

[3] R. Soricut and D. Marcu, "Abstractive headline generation using WIDL-expressions", in Information Processing and Management: an International Journal, vol. 43 no. 6, November 2007, pp. 1536-1548.

[4] M. Banko, V. Mittal, and M. Witbrock, " Headline generation based on statistical translation", in Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000), Hong Kong, 2000, pp. 318–325.

[5] S. Malhotra, and A. Dixit, "An Effective Approach for News Article Summarization", in International Journal of Computer Applications (0975 – 8887) Volume 75– No.17, August 2013.

[6] U. N. Shrawankar, and K.B. Wankhede, "Construction of News Headline from Detailed News Article", in Proceedings of the 10th INDIACom, 3rd International Conference on Computing for Sustainable Global Development, New Delhi (INDIA), IEEE, 16th – 18th March, 2016,[in process].

[7] R. Wang, N. Stokes, W. Doran, J. Dunnion, and J. Carthy, "A Hybrid Statistical/Linguistic Approach To Headline Generation", in Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, IEEE, August 2005, pp 18-21.

[8] T. Bohne, S. Rönnau, and U. M. Borghoff , "Efficient keyword extraction for meaningful document perception", DocEng '11 Proceedings of the 11th ACM symposium on Document engineering, 2011.

[9] I. H. Witten , G. W. Paynter , E. Frank , C. Gutwin , and C. G. Nevill-Manning, "KEA: practical automatic keyphrase extraction", Proceedings of the fourth ACM conference on Digital libraries, Berkeley, California, USA, August 11-14, 1999, pp.254-255.

[10] N. Kumar ,and K. Srinathan, "Automatic keyphrase extraction from scientific documents using N-gram filtration technique", in Proceedings of the eighth ACM symposium on Document engineering, Sao Paulo, Brazil, September 16-19, 2008.

[11] Z. Li, B. He, and Yangnan, "Adding Lexical Chain to Keyphrase Extraction", in 11th Web Information System and Application Conference, 2014.

[12] S. Xu, S. Yang and Fransis C. M. Lau, " Keyword extraction and headline generation using novel work features", in Proceedings of AAAI 2010, pages 1461–1466

[13] S. Siddiqi , and A. Sharan, "Keyword and Keyphrase Extraction from Single Hindi Document using Statistical Approach", in IEEE 2nd International Conference on Signal Processing and Integrated Networks (SPIN) , 2015, pp.713-718.

[14] S. Siddiqi , and A. Sharan, "Keyword and Keyphrase Extraction Techniques: A Literature Review", in International Journal of Computer Applications (0975 – 8887) Volume 109 – No. 2, January 2015.

[15] A. Gupta, A. Dixit, and A. K. Sharma, "A novel statistical and linguistic features based technique for keyword extraction", in IEEE International Conference on Information Systems and Computer Networks (ISCON) , 2014, pp.55-59.

[16] B. Dorr, D. Zajic, and R. Schwartz, "Hedge Trimmer: A Parse-and-Trim Approach to Headline Generation". in Proceedings of the Document Understanding Conference (DUC), 2003

[17] S. H. Atrey, T. V. Prasad, and G. Ram Krishna, "Issues in parsing and POS tagging of Hybrid Language", in IEEE International Conference on Computational Intelligence and Cybernetics, 2012, pp.20-24.

[18] K. Knight and D. Marcu, "Statistics-Based Summarization - Step One: Sentence Compression",in Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, July 30-August 03, 2000, pp.703-710.

[19] B. Dorr, and D. Zajic, "Cross-language headline generation for Hindi", ACM Transaction on Asian Language Information Processing, Vol. 2, No. 3, 2003, pp. 270 – 289.

[20] R. Sun, Y. Zhang, M. Zhang and D. Ji, "Event-Driven Headline Generation", in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, July 26-31, 2015, pages 462–472.

ISBN-13:978-93-5260-126-4, ISBN-10:93-5260-126-2