ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/301221382

Construction of News Headline from Detailed News Article

Conference Paper · March 2016

CITATIONS
7

READS
1,718

2 authors:

Urmila Shrawankar
Rashtrasant Tukadoji Maharaj Nagpur University
223 PUBLICATIONS **780** CITATIONS
SEE PROFILE

Kranti Wankhede
Raisoni Group of Institutions
3 PUBLICATIONS **8** CITATIONS
SEE PROFILE

All content following this page was uploaded by Urmila Shrawankar on 12 April 2016.

The user has requested enhancement of the downloaded file.

Construction of News Headline from Detailed News Article

Urmila Shrawankar
IEEE Member, G.H.Raisoni College of Engineering
Nagpur, MS, India
Email Id: urmila@ieee.org

Kranti Wankhede
Student Researcher, G.H. Raisoni College of Engineering
Nagpur, MS, India
Email Id: w.kranti666@gmail.com

***Abstract* – Usually long news article contains large amount of information. Many a times due to lack of time people are unable to read whole news article. Therefore, headline is required in order to get complete idea of news without reading whole news article. The extractive and abstractive approaches are conventionally used for news headline generation. In this paper, a combinational approach is used for headline construction by using keywords/keyphrases along with parsing technique of Natural Language Processing (NLP). The Keyphrase Extraction Algorithm (KEA) is used to extract keyphrases from input news text. Respective news domain word thesaurus and some other approaches are used for retrieving keywords from news text. Proper headline syntax can be constructed by using parsing technique. Headline is useful to reduce the reading and interpretation time for getting the complete idea of entire news article. The objective is to save reader's time and effort in finding the useful information in a detail news article.**

Keywords – Headline Construction; Keyphrase Extraction Algorithm (KEA); Keyword Extraction ; Natural Language Processing (NLP); Parse Tree Generation.

I. INTRODUCTION

Nowadays the readers do not have enough time to devote for reading the whole news article, so headline cater to their needs. Long news articles generally have longer sentences which contains more words than average length. Headline helps to identify an idea of news article in an optimized way. It also reduces the interpretation time and efforts of reading the whole article.

Headline construction includes news text analysis, understanding the key concept of news and then constructing the headline which reflects vital information of news. Reading each detailed news article is time consuming. The headline is used for finding out important contents of news. The effective headline can be constructed using keywords/keyphrases which are extracted by using keyword extraction methods [1] from news article. According to [2], headline generation is an important tool for reduction of information overload and extracting important points of text. Many techniques are available for improving the quality of news headline [3, 4].

The extractive and abstractive approaches are mainly used for news headline generation. In extractive approach [5], the important sentences from input text are selected and compressed into shortest form. Abstractive approach [4] is also used for headline construction in many researches which tries to understand important contents of news text and then express them in clear and simple language.

In this approach, News headline is constructed by combination of keyword/keyphrase extraction and parsing technique of NLP [6]. Only English news is considered in this work for headline construction. Standard dataset of BBC news of various categories like sports, business, politics, crime and education is used in this work. Keywords/keyphrases which contain important content of document are used for proper headline construction. Numerous methods have been proposed for keyword/keyphrase extraction from text data [1]. Out of those available keyphrase extraction techniques, KEA [7, 8, 9, 10] is used for keyphrase extraction which is used in headline construction. Word dictionary of specific news domain and various other approaches are used as criteria for keyword extraction. Parsing technique of core NLP [6] is used to generate parse tree of leading informative sentences in news article. Generated parse tree helps to construct headline by applying sentence compression algorithm on leading sentences.

The rest of this paper is organized as follows. A comparative study of available headline generation techniques is being presented in section II. Section III portrays the system model for headline construction. The analysis of techniques is explained in section IV. Section V describes the expected result of the system and section VI describes the conclusion of the system.

II. COMPARATIVE STUDY OF HEADLINE GENERATION TECHNIQUES

There are various techniques available to construct news headline. Out of which mostly news headline can be formed by using extractive and abstractive headline generation techniques [3] [5].

In extractive approach, the headline is formed by using sentence compression on informative sentence up to desired length of headline usually 10 words. The compression algorithm uses either symbolic, rule based approach [11], [12] or stochastic, supervised approach [13] for lead sentence compression. The symbolic approach [11] to sentence compression operates on syntactic structure of sentence which usually generates correct and grammatical sentences. But in this system the sentence shrinks too much and sometimes important content information might be lost. To overcome this problem later compressed sentence was padded with important contents of words or phrases. To achieve this, an attempt is made by author [14] to find important contents by computing noun phrase conference chains across input documents or sometimes by discovering topics by Unsupervised Topic Discovery (UTD) algorithm [15]. Both HMM Hedge approach and Hedge Trimmer approach are used by author [16] for cross language headline generation for Hindi documents. Then another attempt is made by author [17] to find the topic keywords from input document using an UTD algorithm. As keywords carry important content information, they are added to compressed sentence which gives increased quality of headline.

In abstractive approach, headlines are created in bottom up manner. The important words and phrases are extracted and glued together to form fluent sentence. A system developed by author [4] uses statistical model for content selection and sentence realization to produce headline. An abstractive headline system created by author [18] used an algorithm first that identifies keywords which are then used to retrieve phrases from input document. Then they are glued together to create headline. The disadvantage of this system is that it is difficult to create fluent output in this manner. Another abstractive headline generation system proposed by author [19] uses statistical model which combines dependency structure and models of n-grams. It starts from dependency structure which is extracted from input document and glues them together using n-grams to smooth the transitions between adjacency dependency structures. But in this method generated headline is not so fluent. Author [3] used WIDL-expressions to generate headlines. Keyword clustering based on several bag-of-words models to construct a headline is used by author [20]. Author [21] worked on event driven model for headline construction.

In this approach the combination of the ideas of previous work is to be used for headline construction. It uses dictionary based approach and some other linguistic features to identify keywords that carry most important content of input news text. As keywords are smallest unit of information, they can provide a compact representation of a document's content [22]. It also uses Keyphrase Extraction Algorithm (KEA) [7] [8] for extracting the keyphrases from input news text. The combination of key words and keyphrase extraction along with parsing technique will help to generate the proper headline construct.

III. SYSTEM MODEL

The System model is divided into five sub-systems as defined below.

- Pre-processing
- Keyword/keyphrase Extraction
- Parse Tree Generation
- Compressed Sentence Generation
- Headline Construction

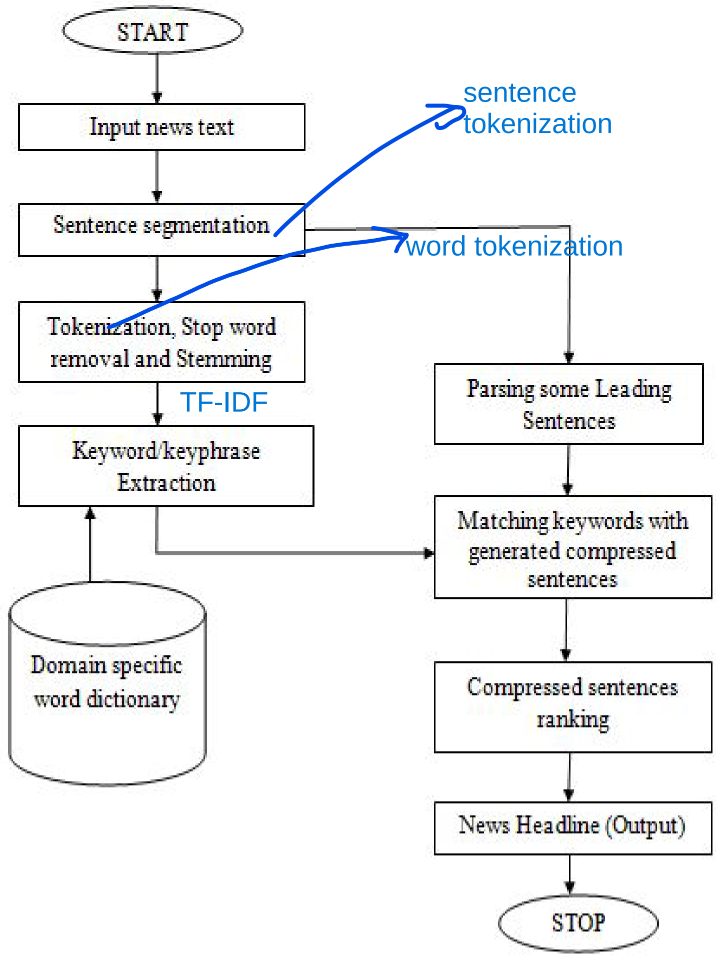
START
Input news text
sentence tokenization
Sentence segmentation
word tokenization
Tokenization, Stop word removal and Stemming
TF-IDF
Keyword/keyphrase Extraction
Domain specific word dictionary
Parsing some Leading Sentences
Matching keywords with generated compressed sentences
Compressed sentences ranking
News Headline (Output)
STOP

Fig. 1. System model.

The steps involved in system model are described in detail as below.

A. Pre-processing

In pre-processing, the following steps can be performed on input text.

1) Sentence segmentation: In this step input paragraph is divided into separate sentences. This can be done by using the criteria of splitting the text by full stop and a space together. The sentence should not terminate at abbreviations by using full stop character e.g. Dr., Ph.D etc.
2) Tokenization: The stream of text is divided into meaningful elements called tokens; like words, symbols, punctuations etc. Then the punctuations are removed from that tokenized text. It is a useful process

in the field of Natural language processing for unique symbol representation.

3) Stop word removal: Prepositions, articles, determiners and directives are removed from the text because it will not contribute any sense to the keywords. The most frequently appearing words are discarded from the input text because these words should not be considered as keywords.
Some of the examples of stop words are: he, in, is, it, a, an, the, and, are, its, of, on, that, the, to, was, were, will, as, at, be, for, from, has, with etc.
4) Stemming: In this step root or base of words is found which helps to avoid repeated counting of same word while extracting keywords.

B. Keyword/Keyphrase Extraction

Keyword/keyphrase extraction is an important phase for constructing the news headline. It is observed that there are lot of approaches for keyword extraction like supervised and unsupervised machine learning, statistical and linguistic methods [23]. In this work, keywords from input news text are extracted using dictionary based approach and some other approaches. In dictionary based approach, the words which are present in dictionary, if found in given input text, can be considered as a keyword and extracted from the text.

The keyword/keyphrase extraction can be done as below.

1) Keyword Extraction: The proposed system identifies the keywords using the subsequent approaches.

a) Term Count: The total number of times a word occurs in a document is known as term count of that word. The vital word is expected to occur many times in a document and hence that word will assign a higher value [23] [24].

b) Term Frequency: Term Frequency (TF) of a word counts the number of occurrences of a word in a document [23]. The TF of a word w in a document d is

$$TF(w,d)=\frac{freq(w,d)}{size(d)} \qquad (1)$$

Where, freq (w, d) is the term frequency of the word w in document d and size (d) is the number of words in the document d [23][24].

c) Temporal Expressions: The words expressing temporal information such as time, week, days, months etc are important in news articles. So they should be extracted as keywords [5].

d) Numeral Data: The numerical values from input news text are extracted as keywords [5].

2) Keyphrase Extraction: In input news text, some phrases are also important for the construction of news headline. Such important phrases are extracted from input text by using Key-phrase extraction algorithm (KEA). Witten in 1999 proposed KEA to extract key-phrases automatically [8].

In this model, KEA performs selection of candidate phrases in three steps:

- Pre-processing is done on input text to tokenize the text and then the punctuation marks, brackets are removed. This process is already completed in previous steps.
- The phrase identification is the second step where KEA considers all subsequences in every line and identifies which key-phrases are suitable. The candidate phrases should not begin or end with meaningless word and it is limited to certain maximum length. word2vec
- The last step includes the grammar patterns which will extract the proper keyphrase as per demand.

C. Parse Tree Generation

The first sentence of lead paragraph of news article contains 86.8% of the headline words [11]. The compression of leading sentence helps for generating title for news stories. In this work, the parse tree of the lead sentences in lead paragraph is generated without affecting the factual correctness or grammar of the sentence. In parsing technique [25], the various phrases present in input sentence is identified and represented in the form of parse tree. The leading sentence of lead paragraph is the input to this stage. The result expected is parse tree generated from lead sentence.

D. Compressed Sentence Generation

The compressed version of some leading sentences of input news text is constructed using compression algorithm [6]. The parse tree generated in previous phase helps to construct compressed version of sentences. The determiners, time expressions, quantifiers and other low content words are removed in the beginning stage of sentence compression algorithm. The larger constituents of parse tree are removed by applying some more compression rules until the required headline length is achieved. The non essential subordinate clauses and relative clauses are also removed by applying various rules found in [11].

E. Headline Construction

The compressed sentences can be obtained from leading paragraph as mentioned above. Then the keywords extracted from input news text are matched in obtained compressed sentences. The compressed sentence having maximum number of keywords is considered as most possible headline and then further processing will be done for proper construction of generated headline. The further processing includes tense identification, grammar checking, syntax correction etc.

IV. ANALYSIS OF TECHNIQUES

The analysis of techniques used for headline construction is as follows.

- Sentence Segmentation- The input news text is divided into sentences according to selected rules for breaking the paragraph into sentences.

- Tokenization-The input news text gets divided into tokens and unwanted punctuations are removed.

- Stop Word Removal-The stop words are removed from input news text so that it should not be considered as keyword.

- Stemming-The porter stemming algorithm is used for stemming of words in input news text in order to avoid the repeated counting of same root word.

- Keyword Extraction-The keywords are extracted from input news text using various approaches like Term count, Term frequency, temporal and numerical data. These extracted keywords help in headline construction.

- Keyphrase Extraction-The Keyphrase Extraction Algorithm (KEA) is used for retrieving the important phrases from input news text. This algorithm produces the list of keyphrases from input text which helps for the construction of headline.

- Parsing of sentences-The parsing technique of NLP generates the parse tree of informative sentences in news text. This generated parse tree helps in formation of compressed sentences.

V. EXPECTED RESULT

Example explained in fig. 2. gives the brief idea of expected result. English news text will be the input to the system. This input news text will undergo various preprocessing steps like sentence segmentation, tokenization, stop word removal and stemming. Then keywords and keyphrases will be extracted from input news text using various approaches which will help to construct the proper news headline. The parsing of some leading sentences will be performed and compressed sentences will generate by applying sentence compression algorithm. Then by matching the extracted keywords in generated compressed sentences, ranking will be done in order to find highest matched sentence. The final output of the system will be the highest ranked compressed sentence which will consider as headline of input news text. Further processing like grammar checking, tense identification etc. will be done in order to get precise headline of detailed news article.

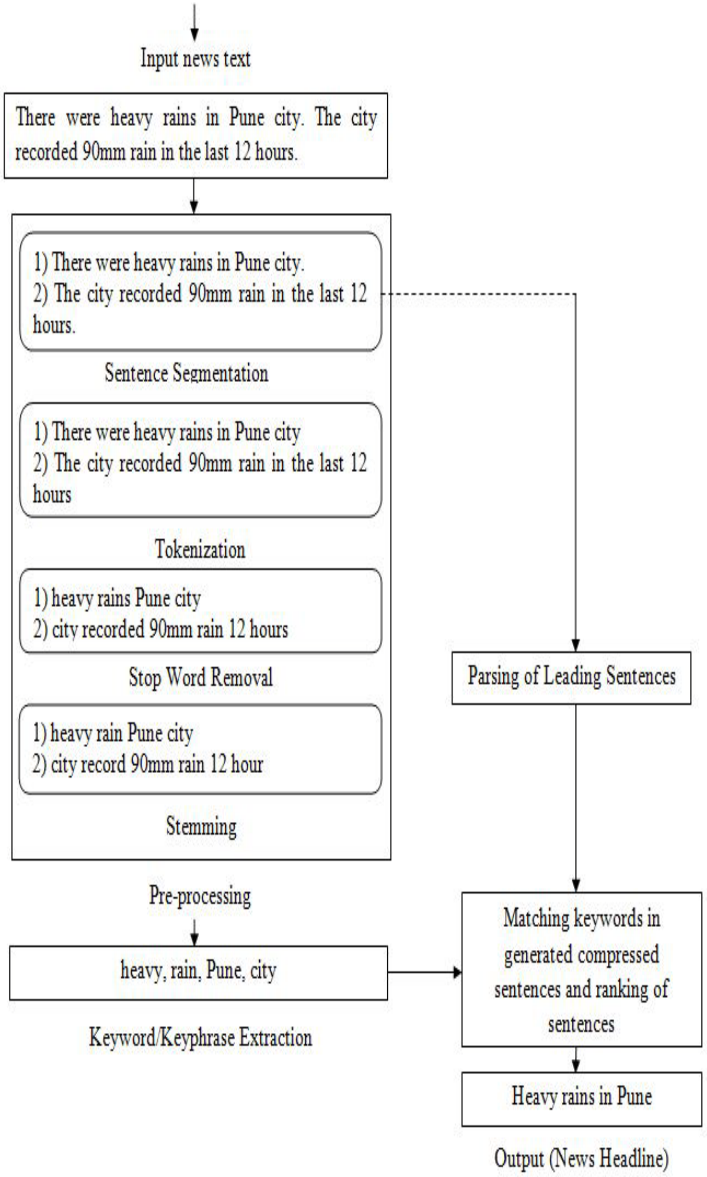
Input news text
There were heavy rains in Pune city. The city recorded 90mm rain in the last 12 hours.
1) There were heavy rains in Pune city.
2) The city recorded 90mm rain in the last 12 hours.
Sentence Segmentation
1) There were heavy rains in Pune city
2) The city recorded 90mm rain in the last 12 hours
Tokenization
1) heavy rains Pune city
2) city recorded 90mm rain 12 hours
Stop Word Removal
1) heavy rain Pune city
2) city record 90mm rain 12 hour
Stemming
Pre-processing
heavy, rain, Pune, city
Keyword/Keyphrase Extraction
Parsing of Leading Sentences
Matching keywords in generated compressed sentences and ranking of sentences
Heavy rains in Pune
Output (News Headline)

Fig. 2. Example of context based news headline construction.

VI. CONCLUSION

This work helps to construct proper news headline by using keyword extraction and parsing technique of core NLP. The linguistic features such as Term count, TF, temporal data etc. and dictionary based approach is used for keyword extraction from input news text. The dictionary used in this system can be updated according to specific domain of input news article. Important keyphrases are extracted by using KEA algorithm which further used for headline construct. The parsing technique of core NLP is used for parse tree generation of most informative sentences of input news text. The sentence compression algorithm is used to generate compressed version of informative sentences from input news text. The keywords are matched in generated compressed sentences and the highest ranked compression sentence is picked out as most possible headline of input news article. Further processing will be performed on generated compressed sentence in order to construct proper news headline. This processing will involve tense correction, grammar checking, proper syntax construction etc.

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, pp.746-759, 2015.

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

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

[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, pp. 318–325, 2000.

[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] 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, pp 18-21, August 2005.

[7] 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.

[8] 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, pp.254-255, August 11-14, 1999 .

[9] 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.

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

[11] 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

[12] R. chandrasekar, C. Doran, and B. Shrinivas, "Motivations and methods for text simplification", *Proceedings of sixteenths international conference on computational linguistics (COLING*), Copenhagen, Denmark, August 05-09, 1996.

[13] 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*, pp.703-710, July 30-August 03, 2000.

[14] S. Burgelar, R. Witte, M. Khalife, Z. Li, and F. Rudzicz, "Using Knowledge poor conference resolution for text summerization", *Proceedings of the document understanding conference (DUC*), Edmonton, Alberta, Canada, 2003.

[15] R. Schwartz, S. Sista, and T. Leek "Unsupervised topic discovery", in *Proceedings of workshop on language modeling and information retrieval*, pp. 72-77, 2001.

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

[17] D. Zajic, B. Dorr and R. Schwartz, "BBN/UMD at DUC-2004: Topiary", in *Proceedings of the Document Understanding Conference (DUC*), 2004.

[18] L. Zhou and E. Hovy, "Headline summarization at ISI", in *Proceedings of document understanding conference(DUC 2003)*, Edmonton, Alberta, Canada , May 31-June 1, 2003.

[19] S. Wan, R. Dale, M. Dras, and C. Paris, "Statistically generated summary sentences: A preliminary evaluation using a dependency relation precision matric", in *Proceeding of workshop on using corpora for natural language generation*, 2005.

[20] 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.

[21] 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*, pages 462–472, Beijing, China, July 26-31, 2015.

[22] 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)* ,pp.713-718, 2015.

[23] 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.

[24] 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)* ,pp.55-59, 2014.

[25] 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*, pp.20-24, 2012.

View publication stats