Middle-East Journal of Scientific Research 23 (6): 1122-1126, 2015
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2015.23.06.22208

Automatic Headline Generation for Tamil Scripts

P. Selvi

Department of Computer science and Engineering,
Tagore Engineering College, Chennai, Tamilnadu, India

Abstract: A headline is recognized as a condensed summary of a document. The necessity for a computerized headline generation has been on the rise because of the need to deal with a wide number of documents, which really is a tedious and time-consuming process. Rather than in place of reading every document, the headline can be utilized to determine those that contain important and relevant information. In this paper, we generate headlines for the Tamil documents. However, the Tamil language includes a different statistical structure compared to the English language and requires special treatment to generate Tamil headlines, especially if we have no dedicated technique for the Tamil language. Headline words are selected using a statistical Translation algorithm which comes under statistical approach. Along the way, headline word selection, different preprocessing steps specific to the language are considered. Grammatical headline formation using a semantic understanding of the language can also be addressed using a clustering algorithm. The generated headlines were assessed using human judgements.

Key words: Headline generation · Summarization · Title creation · Tamil Title Generation

INTRODUCTION

Headline generation is the issue of generating an extremely short summary of a file, which condenses the key ideas discussed in it. Headlines are commonly connected with news articles, but application aspects of headline generation vary from generating table of contents for a file to providing support for interactive query refinement browsing engines. Automatic headlining generation tries to automate the method of providing more relevant or reflective insight in to the input text as opposed to producing likable lines.Generating a headline for a textual document requires analyzing the document, understanding the key concept of the document and making a headline that reflects the information of the document. Therefore, the issue of headline generation concerns complex language processing. A headline is really a condensed summary of a file that accurately represents the key idea of the document. Using this definition,it's obvious that headline generation is really a compressed version of summarization and thus the analysis of headline generation is really a area of the summarization field. The increased number of information emerging in the present day digital world has generated an information overload [1-3]. Information overload describes the issue in understanding a topic and making decisions due to the presence of an excessive amount of information. Therefore, the necessity of automatic headline generation has been raised because of the have to manually review huge amounts of documents, which really is a tedious and time-consuming process. Rather than reading every document, the headline can be utilized to determine which of these contains important or relevant information.

The content in Tamil language on web has already been increasing. There has been many portals, which host large levels of the Tamil language content. However, we argue that the informationhas been underutilized because of the unavailability of language text mining methods in Tamil. Large amount of historical novels, information regarding traditional medicines, historical information can be found in the internet and amount of users has increseing day to day [1-9].

Very little work has been done in languages text processing in Tamil.Accessibility to accessing tools and methods also with a lack of Tamil language [1,9]. So the aim of this research work is to create an effort to create an algorithm to extract keywords from Tamil languages. This will make simple to generate the headline without reading the whole text passage.

Corresponding Author: P. Selvi, Department of Computer science and Engineering, Tagore Engineering College, Chennai, Tamilnadu, India.

Related Work: There are many systems that automatically generate headlines for documents in languages besides Tamil. Automatic head line generation might be classified according numerous dimensions, such as for example like, linguistic versus statistical or extractive versus obstructive. In the extractive approaches [5,6], the most suitable text unit is extracted from the original document and then it may be trimmed to the correct size.A some of the extractive approaches are [10] However, within an abstractive [1-4] headline generation, the original document is analyzed and proper headline words are selected and ordered to represent a consistent and readable headline. On one other hand, statistical approaches [8,9] are on the foundation of the training variation between headline words and the document words from training corpus. Numerous the training approaches are Naïve Bayesian approach [11], a Reverse information retrieval approach. But These methods can be used in cross-lingual headline generation. On one other hand, the linguistic approaches include the usage of information about the structure of the language itself to analyze the document and generate the headline. Linguistically motivated heuristic rules also ideal for generating headlines [7].

Various applications of headline extraction include i) Generation of table of contents for certain document ii) Generation of headlines for every single item in the hit list obtained by web search iii) Text Compression On a tool with limited display or limited bandwidth, the headlines could be a replacement for the entire text- Short message service in Mobile, iv) Voice based application v) Interactive Query Refinement.

Tamil Language: Tamil language is one amongst the Dravidian language, mainly spoken predominantly by Tamil folks of the Indian subcontinent. Tamil words have much more derivational forms than English words. A Tamil word contains constituent parts: a stem, which can be considered as responsible for the nuclear meaning of the verb, attached to which can be zero or perhaps more derivational prefix and zero or one suffix, which together form a word. Tam is really a morphologically rich language causing its relatively high inflectional forms. Normally,nearly all the Tamil words has multiple morphological suffix. The total amount of suffixes is which range between 3 to 13.

The content of Tamil language on the net had been increasing.There were many portals, which host large quantity of the Tamil language content. However, we argue that the data has been underutilized due to the unavailability of text mining methods in Tamil language.

Almost no work has been done for the reason that language processing.So desire to with this research work is to create an attempt to generate the headline from one of the Dravidian languages of Tamil [1,9].

Architecture Framework of Proposed System: The documents are preprocessed before training such as for instance normalization, segmentation, Tokenization and removing stop words and Tagging are performed with the help of Tamil wordnet database. The specific content words called keywords are extracted from preprocessing. Then STHG algorithm is executed in the next phase and generate the headline.

Proposed Algorithm for Headline Generation: The key concept of the used method is always to extract probably the most appropriate group of consecutive words (phrase) from a report body which should represent a satisfactory headline for the document. Each document is represented as a term frequency vector.

The Algorithm for Generate the Headlines Are:

- Transform the documents to a set of terms after words stemming operations.
- Remove stop words. Stop words are normal words which contain no semantic content.
- Vocabulary words are stored in the proper execution of matrix called Vector space model
- Term frequency, Inverse document frequency and TF*IDF for every single word is calculated
- Find the vocabulary words by fixing threshold value for TF*IDF.
- Find the sentence with maximum amount of words having higher TF*IDF value.
- Select the Headline sentence predicated on step 6.

The TF/IDF weighting scheme assigned to the term (t) in document (d) is given by

$$f(t,d) = \frac{f_d(\mathrm{t})}{\max\limits_{\omega \in d} f_d(w)} \quad (1)$$

Where T-term, d-document $f_d(t)$-no of times t appears in document d.

Inter-document characterization is obtained by the idf factor, i.e. the inverse document frequency. Terms that can be found in many documents are less ideal for describing a relevant document from the non-relevant one. The normalized idf factor is computed in the next way:

$$idf(t, D) = \log\left(\frac{|D|}{|\{d \in D: t \in d\}|}\right) \quad (2)$$

The equation (2) could be the inverse document frequency which really is a way of measuring the overall significance of the word t, where is the sum total. amount of documents in the corpus and is the amount of documents in which the term t appears.

The best-known term-weighting schemes use weights which can be written by:

$$w_{i,j} = tf_{i,j} \times idf_i \quad (3)$$

Let V = $\{t_1, t_2 ... t_n\}$ be the list of terms after the stop words removal and words stemming operations. The term frequency vector X_j of document d_j is defined as $X_i = [x_{1i}, x_{2i}, ..., x_{mi}]$.

$$x_{i,j} = \left(\frac{tf}{idf}\right)_{i,j} \quad (4)$$

Using n documents from the corpus, we construct a $m \times n$ term-document matrix X.

Preprocessing Techniques: Before Keyword extraction, the document should really be preprocessed, which means the unstructured data should really be transformed into structured data. Generally, the duty of text preprocessing could be split into four stages namely:

This is actually the example data employed for experimentation is follows.

Normalization: Normalization is the method of removing unnecessary text from the document. The unnecessary text may include text found between hyphens, after bullets and between parentheses.

Segmentation: This phase outputs the document as a couple of words by eliminating the unnecessary semicolons, colons, exclamation marks etc. The full-stops are retained to point end of sentence which will be needed during further stages of headline generation [12].

Parsing: The Tamil shallow parser provides the analysis of a Tamil sentence at various levels. The analysis begins at the morphological level and accumulates at outcomes of POS tagger and chunker. The ultimate ouput combines the outcomes of most these levels and shows them within a representation.

After applying the parsing, following could be the output

Stop Word Removal and Stemming: Stop words are the most popular words in the document which themselves alone doesn't infer any meaning to the sentence.Hence it is needed to eliminate them from the document and headline. Normally, except nouns, verbs, adjectives and adverbs are believed a stop words. So, all the words that are not nouns, verbs, adjectives and adverbs are eliminated.Here Stop words are QC and QF, PSP.Vigneshwaran tamil stemmer has been employed for stemming the most popular root type of words.

Stemming: The segmented words are then queried in the WordNet to obtain the basis type of the words. As an example words like the words "□□□□□□□□□is transformed to □□□□□.The section of speech information, already assigned to each word by the tagger, is employed here to query the WordNet to obtain the basis type of the term in the sense by which it is found in the document [13].

Tamil Wordnet: The Tamil Wordnet is alexical-semantic network that's structured along exactly the same lines since thePrinceton WordNet (lexical reference system). In a wordnet, which will be basically a semantic network, the various lexical kinds of words as nouns, verbs etc. are organised into'synsets'called sets of synonyms. Each synset represents a lexical concept and they could be linked by several types of relation such as for example hyperonymy, antonymy, etc. For obvious reasons, all wordnets resort to the exact same system of synset identification.To get the synonym matching,the Tamilwordnet has been used.In the example,the words "□□□□□□□□ □□□□□□□ and □□□□□" refers exactly the same meaning.

Training Corpus: We have collected a comprehensive data set consisting of articles with human generated headlines. The next would be the categories:

- Business Articles
- Business news articles.
- General articles
- News Articles

The experiments has been performed on 5000 documents of the FIRE (Forum of Information Retrieval Evaluation) Tamil corpus extracted from newspapers such as for example Dinamani and Dinamalar [13].

The news articles from various news sites has been downloaded. The headlines texts were downloaded based on a snapshot of the links within main page of news sites

like Dinamalar, Dinamani. There are around 20000 news articles making use of their respective headlines. Since headlines generated by our system are typical words/phrases extracted from the body of the articles, we reduced the set to only around 15000 articles, each which contains most of its headline words.

RESULTS

The application form of the keyword extraction is illustrated having an example. Given text segments, as shown below in Figure 4. The Table 1 shows the score that reflects their significance of the keywords in the paragraph.The minimum threshold value has been set for evaluation is 0.1.The values above 0.1 have considered for calculation.Following the finding words which are essential for generate the headline. The positioning of the term should be identified.For each and every line, the most quantity of words reached the threshold value has been calculated. In that, which sentence having higher quantity of words, that sentence would be selected as the headline for the given passage. In the example, the initial sentence having four words out of five words i.e. 80% which is higher percentage among other sentence.Hence, the first sentence will be selected as the headline.

Table 2 shows the scores on the test data. The proposed approach measures frequency of unigrams which cross the threshold value presented in each sentence.From the Table 2.

It is clear that the first sentence scores the highest result in the automatically generated headlines (Figure 4). Unfortunately, there are no available results on an Tamil language headline generation system to compare with and it is not right to compare these results with other systems applied to other languages or different data sets.

Evaluation: We did an informal evaluation of the headlines generated for 40 tamil documents from FIR and 20 from Dinamalar news paper. In FIR articles is becomes successful in generating headlines 70% of the case. Some cases in generate totally absurd headline or no headline at all. These could be improved using large amount of training data with extractive headline.There's improvement of the end result in case there is Dinamalar newspaper. Within our experiments we tried to perform our algorithm for different group of values of position penalty, string penalty and gap penalty. There parameter could be study from working out data and bears further study.

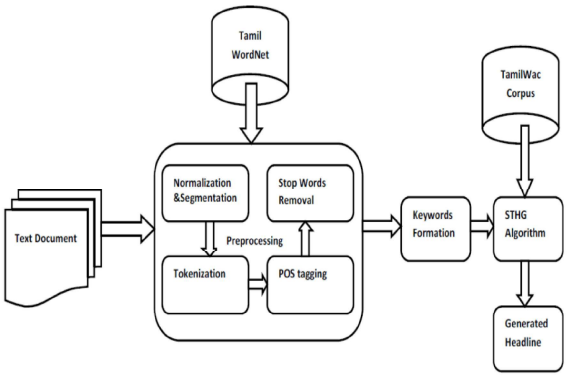

Fig. 1: Architecture of Automatic Headline Genaration System

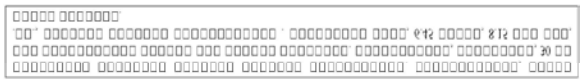

Fig. 2: Sample data

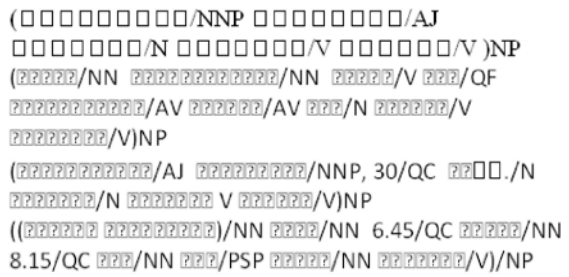

Fig. 3: Result after Parsing

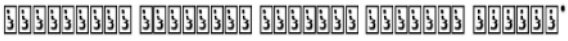

Fig. 4: Generated Headline

Table 1: Result of TF/IDF calculation

Words	No of Terms	TF	IDF	TF*IDF
□□□□□□□□□/ □□□□□	3	0.20	0.70	0.14
□□□□□□□□ □□□□□□□/ □□□□□	2	0.13	0.88	0.12
□□□	4	0.27	0.57	0.15
□□□□□□□	2	0.13	0.88	0.12
□□□□□□	1	0.07	1.18	0.08
□□□□□□□□□□□□	1	0.07	1.18	0.08
□□□□□	1	0.07	1.18	0.08
□□□□□□□□□□□	1	0.07	1.18	0.08
□□□□□□	1	0.07	1.18	0.08
□□□□□□	2	0.13	0.88	0.12
□□□□□□□□	1	0.07	1.18	0.08
□□□□□□ □□□□□□□□□	1	0.07	1.18	0.08
□□□□	1	0.07	1.18	0.08

Table 2: Selection of main sentence.

Sentence	key words in (%)
[illegible]	80
[illegible]	25
[illegible]	16
[illegible]	11

CONCLUSION

The STHG algorithm using TF-IDF feature selection is pro-posed for generating headline in Tamil languages.The significance of the term is on the basis of the value of TF*IDF. The amount of keywords selection is on the basis of the threshold value TF*IDF. Precision and recall have a tendency to antagonize each other. Which means that efforts to improve precision will generally compromise recall and efforts to improve recall will generally compromise precision.

As we've observed in the graph there's a steep rise in the recall. The future development is to enhance the end result by eliminating the steep increase to gradual increase by taking into consideration the features of the text.

This algorithm shows satisfactory results for test data in Tamil language.The outcomes could be improved by considering vocabulary pruning methods like lemmatization and stemming.

REFERENCES

1. Vijayapalreddy P. Dr. B. Vishnu Vardhan, Dr. A. Govardhan and M. Yesy babu, 2011. Statistical translation based headline generation sytem for Telugu, IJCSNS International Journal of Computer Science and Network Security, 11(6).
2. Alotaiby Fahad, Salah Foda and Ibrahim Alkharashi, 2012. New approaches to automatic headline generation for Arabic documents, Journal of Engineering and Computer Innovations, 3(1): 11-25.
3. Alotaiby, Fahad, A., Automatic Headline Generation using Character Cross-Correlation, Proceedings of the ACL-HLT 2011 Student Session, pp: 117-121.
4. Wang Ruichao, John Dunnion and Joe Carthy, Machine Learning Approach To Augmenting News Headline Generation, In Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)-2005
5. Mondal Amit Kumar and Dipak Kumar Maji, Improved Algorithms For Keyword Extraction and Headline Generation From Unstructured Text.
6. Keyword Extraction and Headline Generation Using Novel Word Features,2010. Conference: Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA.
7. Bonnie Dorr, David Zajic, Richard Schwartz, Hedge Trimmer: A Parse-and-Trim Approach to Headline Generation-HLT-NAACL-DUC'03 Proceedings of the HLT-NAACL 03 on Text summarization workshop, 5: 1-8.
8. Jin, R. and A.G. Hauptmann, 2000. Headline generation using a training corpus. In 2nd International Conference on Intelligent Text Processing and Computational Linguistics, CICLING.
9. Hanumanthappa, M., M. Narayana Swamy and N.M. Jyothi, 2014. Automatic Keyword Extraction from Dravidian Language, IJISET - International Journal of Innovative Science, Engineering & Technology, 1(8).
10. Songhua Xu, Shaohui Yang and Francis C.M. Lau, Keyword Extraction and Headline Generation Using NovelWord Features, Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10).
11. Alfonseca Enrique, Jose Mari, Antonio Moreno-sandoval, 2004. A Study of Chunk-Based and Keyword-Based Approaches for Generating Headlines, TAL - Natural Language Processing - TAL, pp: 395-406.
12. Thangarasu, M. and Dr. R. Manavalan, 2013. Stemmers for Tamil Language: Performance Analyses, International Journal for Computer Science & Engineering Technology, 4(7): 902-908.
13. Tamil Corpus Data, Forum for Information Retieval Evaluation(FIRE). http://irsi.res.in/fire/data.html.