ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

A Study of Natural Language Processing Based Algorithms for Text Summarization

Prashant G Desai[1], Saroja Devi H[2], Niranjan N Chiplunkar[3] , Mahesh Kini M[4]

Research Scholar, Department of Computer Science and Engineering, NMAMIT, Nitte, Karnataka, India[1]

Professor and Head, Department of Computer Science and Engineering, NMAMIT, Nitte, Karnataka, India[2]

Principal, NMAMIT, Nitte, Karnataka, India[3]

Assistant Professor, Department of Computer Science and Engineering, NMAMIT, Nitte, Karnataka, India[4]

ABSTRACT: Internet is the largest housing for the information. This information is stored in different formats such as HTML, XML, Word, pdf documents. The challenge with the natural language text is that there is no specific way in which the text is arranged. Therefore, in this context, the text mining approach gains significance. The concept of text mining is used for extracting meaningful and useful information from the natural language text.

KEYWORDS: Classification, Summarization, NLP, Template, Evaluation

I. INTRODUCTION

In the era of ever growth of digital information, the Natural Language Processing (NLP) has become a significant area of research. NLP is an approach to analyze the text with the assistance of computer, to achieve human-like language understanding [10]. The research in the field of NLP has been concentrating on the concepts such as Text Classification, Automatic Text Summarization, Events extraction, Dialogue Management, clustering, topic discovery to mention a few. Present research concentrates on classifying the relevant documents from the pool of documents using Naïve-Bayes algorithm, Automatic Text Summarization and dialogue management.

Text classification can be defined the process of assigning similar documents to the same category. Here machine learning algorithm is trained with dataset to build a classifier. This classifier is then used to predict the category of new data.

Automatic text summarization is the task of reducing large amount text present inside a document to a high quality, useful and meaningful small amount of text. There are two types of text summarization methods: Abstractive Text Summarization and Extractive Text Summarization

Question answering can be defined as the task of extracting the answer against the question presented in natural language text [2, 16].

II. RELATED WORK

A frequent term based text summarization method with the help of HMM tagger is presented in [6]. The summarization is done with help of concepts case folding, tokenization, stop word removal, stemming, feature term identification, POS tagging, noun and verb chunking, term frequency, term weight and then sentence scoring analysis.

A summarization technique called SUMMARIST is discussed in [3]. Apart from the regular pre-processing of the natural language text, it uses additional concepts topic identification with the help of cue phrases and topic interpretation to arrive at the summary.

The approach adopted by the researchers of [9] concentrates on extracting key phrases, selection of sentences having key phrases within them. The similarity measure is used to check the relation between candidate sentences and title of the document. Finally the sentences of highest ranks are chosen as the part of summary.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

An algorithm using generic relevance weight based on non-negative matrix factorization (NMF) is reported in [13]. The summary is generated based on the sentences containing major and subtopics of the search results. The important feature of the methodology used for summarization in [4] is the use of fuzzy logic. That is, a value from zero to one is obtained for each sentence in the output based on sentence features and the available rules in the knowledge base. The obtained value in the output determines the degree of the importance of the sentence in the final summary.
Machine learning algorithm that uses both generic and query based summaries has been proposed by the authors of [5]. Different features of the sentences are automatically combined to rank the sentences. Then words occurring in the same sentences are grouped into word-clusters to be used as features.
Researchers investigate a two-step sentence extraction method for text summarization [15]. In the first step, two adjacent sentences are joined to create bigram psudo. Then statistical methods based on title and location features are used for calculating the importance of those sentences. Then in the second step bigram pseudo sentences are separated into their original sentences and second sentence extraction is done by adding aggregation similarity method. The aggregation similarity method finds the significance of the sentence by calculating the similarities of all other sentences in a document. It offers a low cost and robust algorithm as there is no use of linguistic resources such as WORDNET.
An unsupervised summarization method that creates the summary by clustering and extracting sentences from the original text document is implemented in [11]. For this purpose new criterion functions for sentence clustering have been proposed. Also researchers have developed a discrete differential evolution algorithm to optimize the criterion functions.

III. OUR APPROACH

Part-of-Speech (POS) tagging is an important step for pre-processing the document. In the initial part of our research [7] we concentrated on building a POS tagger. The idea was to build POS tagger using open source softwares and then use the same while preprocessing the documents. The work was implemented successfully. However, the serious limitation was that the vocabulary used for building the same was not comprehensive. Therefore, later it was decided to switch over to Stanford POS tagger [12] for the purpose of POS tagging.

The work presented in the Prashant G Desai et. al. [8], algorithms for automatic text summarization and dialogue management were implemented. The methodology of automatic text summarization algorithm is based on the requirements of user before original text is summarized while the algorithm used for dialogue management establishes interactions between user and the computer. The main contribution of work reported here is the template based algorithm for text summarization and dialogue management. The algorithm allows the user to specify the requirements and type of summary to be generated. We call this facilitation a template. Then the algorithm works on the template as provided by the user to generate the summary. The experiments conducted during the implementation of work have produced encouraging results.

IV. EVALUATION AND ANALYSIS OF RESULTS

There are different methods available for evaluating the summaries. [1] presents two methods: intrinsic and extrinsic. In intrinsic summary evaluation, different metrics used to compare the automatically generated summary with reference summary generated by the human beings. Certain factors used in intrinsic summary evaluation are calculating precision-recall-F-Measure, similarity measures using Cosine similarity, Jaccard Coefficient, Euclidean Distance and many more. Extrinsic evaluation is based on acceptability, accuracy and how it affects on tasks like judging relevance or comprehension reading [14]. Majority of the researchers use intrinsic summary evaluation techniques based on calculation of Precision, recall and F-measure. However, we have adopted intrinsic evaluation based on Cosine Similarity since it compares automatically generated summary with a manually (Reference) generated summary [8]. Table -1 represents study of the performance of algorithms used in automatic text summarization proposed in present work [Sl.No. 1] and various algorithms proposed by previous researchers. It shows that the methodology used in the present research achieves better precisions, which is found to be higher than several previously proposed algorithms.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

Table 1: Key Features and Performance of Various Algorithms Used in Text Summarization

Sl. No.	Algorithm	Key Features	Evaluation Method	Results
1	Template Based Algorithm	Allows user to prepare template that has provisions to specify events, locations, named Entities. User has the provision for specifying any number of and any type of POS patterns	Cosine Similarity	Average Similarity of Automated Summary with manual summary is 71.80%
2	Automated Text Summarization in SUMMARIST	modules for position, word frequency, cue phrases, Topic interpretation by 2 or more topics fusion	Precision, Recall metrics are used for evaluation	Overall Precision 69.31 %
3	Automatic Text Summarization Based on Word-Clusters and Ranking Algorithms	Uses ranking features such as Title key word, Local Context Expansion with the help of WordNet, term frequency Acronyms, Cue Words, Common terms	Precision, Recall metrics are used for evaluation	Overall Precision 70 %
4	Automatic Text Summarization Using Two-Step Sentence Extraction	combines two adjacent sentences into bi-gram pseudo-sentence and removes unwanted data. Statistical methods such as title, location, frequency, aggregations are adopted. Separate the combined sentence and extract the second sentence.	F1 measure based on precision and Recall metrics is used	Overall Precision 51 %
5	Corpus based Automatic Text Summarization System with HMM Tagger	Feature term identification, HMM based POS tagging, Noun-Verb chunking, term frequency and term weight , sentence scoring	Not Evaluated	Not Evaluated

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

6	Automatic Personalized Text Summarization Agent using Generic Relevance Weight based on NMF	Sentence ranking, Generic Relevance weight. That is extracting sentences covering the major and sub topics of the search results with respect to user interest.	Precision, Recall and F-Measure metrics are used	Overall Precision 40 %
7	Text Summarization Extraction System Using Extracted Keywords	Term frequency, Inverse Document Frequency, document title and font type, Limited number of POS patterns approach	Precision, Recall metrics are used for evaluation	Overall Precision 70 %
8	Feature-Based Sentence Extraction Using Fuzzy Inference rules	Features like Title feature, Sentence length, Term weight, Sentence position, Thematic word, etc and the fuzzy logic is applied	Precision, Recall metrics are used for evaluation	Average precision of 44.82%

V. CONCLUSION

Processing and understanding natural language text and automatic summarization has been a challenging research area since many years. It is being applied in the domains such as research documents, opinion mining, product reviews, education domains, emails and blogs etc. However, a particular summarizer is not best suited to all the domains. Most of the techniques relied upon features such as title, position of the sentence, sentence length. Our work concentrated on formation of rules according to the requirements of end users in the form of a template. The experiments conducted and results obtained are satisfying.

REFERENCES

[1] Ahmed A. Mohamed, Sanguthevar Rajasekaran, "A Text Summarizer Based on Meta-Search", Proceedings of IEEE International Symposium on Signal Processing and Information Technology, pp.670-674, 2005
[2] Daniel Sonntag, "Distributed NLP and Machine Learning for Question Answering Grid", Proceedings of ECAI, 2004
[3]Eduard Hovy, Chin-Yew Lin, "Automated Text Summarization in SUMMARIST", Proceedings of Advances in Automated Text Summarization, pp.2-14, 1999
[4] Ladda Suanmali, Naomie Salim, Mohammed Salem Binwahlan, "Feature-Based Sentence Extraction Using Fuzzy Inference rules", Proceedings of International Conference on Signal Processing Systems, IEEE, pp.511-515, 2009
[5] Massih R. Amini, Nicolas Usunier, Patrick Gallinari, "Automatic Text Summarization Based on Word-Clusters and Ranking Algorithms", Proceedings of ECIR, pp.142-156, 2005
[6] M.Suneetha, S. Sameen Fatima, "Corpus based Automatic Text Summarization System with HMM Tagger", Proceedings of International Journal of Soft Computing and Engineering, ISSN: 2231-2307, Vol. 1, Issue-3, pp.118-123, 2011
[7] Prashant G. Desai , Niranjan N. Chiplumkar, Saroja Devi, "An Approach for POS Tagging for a Conversational Software System", Proceedings of the First International Conference on Research Trends in Computer Technologies, pp.423-425, 2013
[8] Prashant G. Desai , Saroja Devi H ,Niranjan N. Chiplumkar, "A Template Based Algorithm for Automatic Summarization and Dialogue Management for Text Documents", Proceedings of International Journal of Research in Engineering and Technology, Vol. 04 Issue: 11, pp334-340, 2015
[9] Rafeeq Al-Hashemi, "Text Summarization Extraction System (TSES) Using Extracted Keywords", Proceedings of International Arab Journal of e-Technology, Vol. 1, No. 4, pp. 164-168, 2010
[10] Raju Barskar, Gulfishan Firdose Ahmed, Nepal Barska, "An Approach for Extracting Exact Answers to Question Answering (QA) System for English Sentences", Proceedings of ICCTSD, pp. 1187 – 1194, 2012

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

[11] Rasim Alguliev, Ramiz Aliguliyev, "Evolutionary Algorithm for Extractive Text Summarization", Proceedings of Intelligent Information Management, pp. 128-138, 2009
[12] Stanford University POS Tagger
[13] Sun Park, "Automatic Personalized Text Summarization Agent using Generic Relevance Weight based on NMF", Proceedings of Honam University Gwangju, pp. 142-144, 2008
[14] Vishal Gupta, "A Survey of Various Summary Evaluation Techniques", Volume 4, Issue 1, pp159-162, 2014
[15] Wooncheol Jung, Youngjoong Ko, Jungyun Seo, "Automatic Text Summarization Using Two-Step Sentence Extraction", Proceedings of AIRS, pp. 71-81, 2005
[16] Zhiguo Gong 1, Mei Pou Chan, "iQA: An Intelligent Question Answering System"