Automatic Extraction of Features from Stories Using Extractive Techniques

[1]Deepali Vaijinath Sawane, [2]Sanah Nashir Sayyed, [3]C. Namrata Mahender

1,2Research Student , [3]Assistant Professor, Dept. of Computer Science & IT , Dr. B. A. M. U., Aurangabad, Maharashtra, India.

[1]deepalisawane99@gmail.com, [2]sanahsayyed92@gmail.com, [3]nam.mah@gmail.com

Abstract - Stories are significant part of life because from childhood to now we are learning vital lesson through story, summarization of story helps to get the core of the story. Human being are emotional creatures, summarization of story allow us to get any information quickly because information is connected with emotion .Story summarization will help us in teaching too, in remembering important concept and relating concept to another concept. Story plays important role in children by enhancing their listening skill and overall growth. Summarization of story is most powerful marketing tool; it grasps attention of customer, if you inform to customer of your product through story. Information extraction from the text have many important application in document verify, detection of vehicles number plates, relation of stories, location, character etc. In this paper more attention is given on verb and noun. Because verb gives idea of important event and noun relates to particular entity from the story. The purpose is to present this paper to give the introduction with some basic information extraction concept, which will be very useful for the reader to get vital information within fraction of second and Text Extraction Features such as Title, actor, moral, location etc. Information extraction can also be used for extract fillers for a predetermined set of slots (roles) in a particular template (frame) relevant to the domain.

Keywords: Text Summarization, Natural Language processing, Extractive summary, Named Entity Recognition, POS Tagging.

I . INTRODUCTION

Information extraction (IE) from text is concerned with extraction the relevant text data from a collection of documents [1]. Information extraction is the process of automatically extraction vital and valuable information entities or the relationship between different information entities from original documents and it presenting in a structured form [2]. In many documents, the text is stored as a set of string objects, each object gather a combination of character, coordination the fonts and other information. Information extraction process structured data from unstructured text by recognizing reference to named entities as well as builds bonds between such entities [3]. Natural language understanding is critical for bulk information extraction task because the desired information can only be identified by conceptual roles [4].

For example, we can take stories, titles, location, relation, characters and personal pronouns. Nowadays, increasing number of document is available in text format as considered as favorite file[5].

The advantage of document analysis is that the character and layout information is obtained from text document. The text extraction feature, text is being extracted from original documents[6]. The text extraction is done with the help of character description and stoke configuration. Firstly from the text will be detected from the original text, understood and then recognized. Information extraction identifies important sentences which are directly selected from document [7]. Information extraction collects important text form original data without changing its meaning. Text Summarization is the method of automatically creating a compressed version of a original text document that helpful for user to collect important information .This information Content a summary as per the requirement of user. This technique presents a method for identifying some feature with the help of Name Entity Recognition [8]. Text Summarization is the technique to highlights important information from original document in few words as possible. Section II Presents concept of information extraction. Section III Presents the literature review, Section IV shows the experiment result, Section V Presents the conclusion

A] Type of summerization:

1] Extractive Summarization

Extractive Summarization extracts important text from the given text and groups them to generate a summary without

DOI : 10.18231/2454-9150.2018.1301

changing text. Usually, sentences are arranged in the similar order like in the main text document [9].

2] **Abstractive summarization**

To examine the text abstractive summarization can be done by understanding the original text with help of linguistic method. The main purpose of abstractive summarization is to develop relevant summary that may be able to show information in an accurate way that commonly needs highly developed language generation techniques [9].

B] Summarization Techniques:

In Addition to Abstraction Text Summarization and Extraction Text Summarization, there are various types of Summarization [10]. There are many different type of Summarization Techniques might be useful in Various application can be categorized based on Approaches, Type of Details, Type of Content, Limitation, Number of Input Document and Languages as follows:

1] Based on Approaches:

The Extractive Text Summarization and Abstractive Text Summarization are two strategies for Summarization. The Extractive Text Summarization consists of source sentences as it is and adding into summary. Abstractive Text Summarization involves generating novel sentences for summary.

2] Based on Type of Details:

The Type of details summary is used for short summary of a lengthy document and it does not change the important idea of original document. Types of details summary is either informative or indicative. It increases the excitement of a user to read original document.

3] Based on Type of Content:

In the Type of Content Summary all information is at level of importance of the original document which is not user specific. The generic summarization system is user friendly and does not depend on original document.

4] Based on Limitation:

In the Summary type of Based on limitation is accept input text as like newspaper, articles, stories and manuals etc. It does not accept the general types of document. It is limited to the some special type of input.

5] Based on number of Input Document:

In the type of Summary Based on Number of Input Document can be either Single Document Summarization or Multi Document Summarization. Single Document Summarization can accept only one document as input and Multi document Summarization accept more than one Input Documents. Single document Summarization is easier to produce Summary of Single Document. Multi Document Summarization are difficult to produce Multi Document summary as Compare to Single Document Summary.

6] Based on Language:

In the Summary Type Based on Languages are classified into

Mono Lingual System and Multi Lingual System. Mono Lingual system accept document of Some particular language. The Multi Lingual System accept Document in Different Languages.

C] How Summarization Plays An Important Role:

Summarization is an important skill for students, teachers, and readers to learn important ideas, projects as like Science, math, mini projects, and major projects. It helps students to learn to determine essential ideas and consolidate important details that support to them. It enables students to focus on keyword phrases of an assigned text that are worth nothing and remembering. Summarization has become the necessity of many applications for example search engine, business analysis, market review. Summarization helps to gain required information in less time.

D] Text Extraction:

An Extractive Text Summarization method consist of selecting important sentences, paragraphs etc, from the original document and concatenating them into shorter from. The importance of sentences is decided based on statistical and linguistic features of sentences. Text extraction system is proposed based on pos tagging by considering Hidden Markova Model using corpus to extract important features to build a summary[11]. Text Extraction identifies and extracts key sentences or words from the source text and concatenates them to form a concise summary.

II. LITERATURE REVIEW

TABLE 1: LITERATURE REVIEW

AUTHOR & YEAR	METHOD/ TECHNIQUE	DESCRIPTION	LACUNA
Rada Mihalcea, Niraj Kumar, Kannan Srinathan and Vasudeva Varma, (2013), Sarda A.T. and Kulkarni A.R.(2015) . [11][12]	Graph Theoretic Approach	This representations passage provides a method of identification of themes. -After the common pre-processing steps, namely, stemming and stop word removal.	There is Limited Linguistic support to include only stop words, stemmer and punctuation marks to achieve the goal.
Khosrow Kaikhan(2004), Sarda A.T. and Kulkarni A.R.(2015) [13][14]	Text summarization With Neural Networks	This method involves training the neural networks to learn the types of sentences that should be included in the summary.	There is most the Article did not have subtitle or section heading.

		Three- layered Feed Forward neural network are used in this method.	
Zhang Pei-ying and Li Cun-he(2009), Mehdi Bazghandi, Ghamarnaz Tadayon Tabrizi and Majid Vafaei Jahan (2012), Anjali R. Deshpande, Lobo L. M. R. J. (2013).[15][16]	Cluster Based Method	-It is intuitive to think that summaries should address different "themes" appearing in the documents. -If the document collection for which summary is being produced is of totally different topics, document clustering becomes almost essential to generate a meaningful summary.	Limitation is All the methods of Summarization is done with the of 50% only.
Ibrahim Imam, Nihal Nounou, Alaa Hamouda, Hebat Allah Abdul Khalek(2013), Ahmed A. Mohamed, Sanguthevar Rajasekaran, (2006), A. P. Siva kumar, Dr. P. Premchand and Dr. A. Govardhan(2011). [17][18]	Query Based Extractive Text Summarization	-In query based text summarization system, the sentences in a given document are scored based on the frequency counts of terms. -It uses Vector Space Model [11].	There is need to improve the system by adding sentence simplification Techniques for simplify the sentence which are very large & Complex.

III. PROPOSED WORK

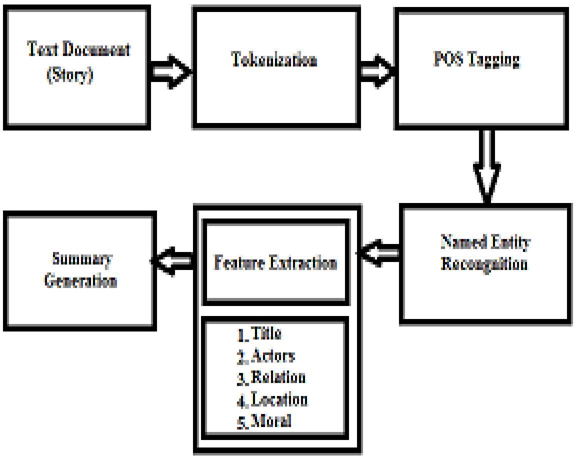
For the Present work Small stories have been considered as raw input, such as S1, S2, S3, S4, S5, S6, S7.S8, S9, and S10.

The Rule Applied of NER and with used of Extractive summarization Technique feature are extracted.

[A] Information Extraction

Information extraction is the method of automatically extraction most of the important information entities or the relationship from text document and presenting it in a specific form. One particular type of information extraction is named entity recognition, it involves identifying references of particular kind of people, company and location like this objects[19]. Information extraction combines technique from natural language processing, lexical resources and semantic constraints. Information Extraction is used to find the attribute of any specific company, any specific stories, any news, employee, article, books, college such this type of object[20]. It is so helpful for user or reader, means if user wants any objects attribute then he can get that attribute with the help of information Extraction System. The method to improve quality of Extractive summary and it works in some phases. The following Figure shows Text Extractive techniques:

Figure 1: The Text Extraction framework



A] Text Documents (Story) :

This is the general Text documents as it available in original form ,Some Type of Stories, News paper, e-book .

B] Tokenization : A term tokenization the process of splitting larger text the text into smaller text part of any General Text document. It is sequence of character in some particular group and is useful for processing.

C] POS Tagging :

A POS(Part of Speech) tagging is also a token , such as Noun, Verb and Adjective. We use POS tagging in our Work for Extracting Noun, Verb and Adjectives from the Stories.

D] Named Entity Recognition:

A NER is used to entity Identification Such as Noun, Proper noun, such as some entity refers type of noun like Monkey NN, Friends NNS, Dove NNP.

E] Feature Extraction :

With using Extractive techniques those words, phrases and feature are plays an important role in that documents or

DOI : 10.18231/2454-9150.2018.1301

Stories are Extracted as like Title, Actors, Relation , Location and Moral.

F] Summary Generation :

All important text as like word , phrases , title, location, relation and moral is extracted by Extractive Technique is displays as Summary.

G] Text Extraction Features:

The Text Extraction System is proposed based on POS tagging by considering Hidden Markov Model using corpus to extract important Features to build a summary .Text Extraction identify and Extract key sentences or words from the source text and concatenate them to form a concise summary.

TABLE 2 : FEATURE EXTRACTION

Features	Description
Title	The First line of the document that occurs in the sentence of the document that contain words or their synonyms in the title should be considered for inclusion in the summary as they reveal the theme of the document.
Actors	For the Text Extraction sentences having proper noun are important like name of person, Animal, organization, place, city etc.
Location	It depends on the Event that occurs on the some particular places, Organization, village etc.
Relation	In Structured data there is a regular and predictable organization of entities and relationships. For example relation between employee and Company etc .
Moral	A lesion that can be derived from a story or experience. There are standard of behaviors and principles of right and wrong.

The five field of information extraction is presented along these five dimensions:
[A] The type of structured extracted.
[B] The type unstructured sources.
[c] The type of Input resource available.
[D] The method used for extraction.
[E] The output of extraction.

Table 3 : Shows The Overall Description Of The Data.

Story Name	Nomenclature used for story names in present work	Actual length of story	Min. Sentences length in story	Max. length of sentence in story
The Ant and The Dove	S1	12	10	15
The Clever Monkey	S2	11	10	13
The Fisherman And The Little Fish	S3	8	7	9
The Foolish Rabbit	S4	6	10	12
The Fox and the Stork	S5	6	7	10
The Hare And The Hound	S6	7	8	11
The Lion And The Boar	S7	8	9	12
The Peacock And The Juno	S8	6	7	11
The Town Mouse and The Country Mouse	S9	6	9	13
who will bell the cat?	S10	7	8	14

Secondly the sentences are POS tagged . The tags are very useful for further processing mainly in identification of unique nouns, verbs from stories. Example of POS tagged sentence :[(, ('neighbors', 'NNS'), ('were', 'VBD'), ('having', 'VBG'), ('trouble', 'NN'), ('with', 'IN'), ('their', 'PRP$'), ('crops', 'NNS') , ('powerful' , 'JJ')]

Named Entity Recognition : The task of identifying proper names of people, organizations, locations, or other entities is a subtask of information extraction from natural language documents.

etc. e.g. ['Raj', 'PERSON'], ['Tortoise', 'ANIMAL'], ['Jungle', 'LOCATION'], ['Friends', 'RELATION'].

A] Extraction of nouns:

In entity Recognition there are extracted some words from the stories, a word other than a pronoun used to identify any Of a class people, places or Common noun or to name of Particular. Those words we can call extracted noun in Table 4.

Table 4: Extracted nouns

Story name	Extracted nouns
S1	Day, Ant, Dove, Water, Time, Spring, blade, Way, Tree, Leaf, Hunter, Trouble, heel, Heel, turn, net, pain, grass, Towards, Begets .
S2	Monkey, Crocodile, river, tree, home, Way, wife, desire, heart, friend, bank, Couple.
S3	Fish, day, catch, livelihood, profit, gain, Wise, money, river, use, sir, kind, Desperation.
S4	Rabbit , lion, nut, inspection, chaos, Jungle, king, sky, way, head.
S5	Stork , Day, prank, table, dish, soup, fox, Meal, time, beak, kindness, fox, dinner, Jar, ate, well, turn.
S6	Story, lesson, day, hound, hare, time, Herd, one, beast, rabbit, life, dinner, Incentive, action, hunt.
S7	lion, boar, summer, day, reach, water, Body, drink, while, breath, food, fight, Advantage, defeat.
S8	Peacock, nightingale, latter, goddess, Voice, beauty, eagle, strength, Everyone, way, everything.
S9	Mouse , town, food, bacon, cousin, Country, day.
S10	cat, mouse, bell, neck, none, lot, enemy, Night, nice, horde, wise.

B]Extraction of verb:

A word used to describe an action, state, or occurrence, and forming the main part of the predicate of a sentence, those

DOI : 10.18231/2454-9150.2018.1301

words are verbs . The verbs extracted from the stories in Table 5.

Table 5 : Extracted Verb

Story name	Extracted verbs
S1	Dry, throw , trap, do, fly, reach, Climb, have.
S2	Taste , understand , take, get, trust, Bring, be, eat, cross, reach, use.
S3	Catch, live, let, grow, make, give, Forgo.
S4	Be, Was, fear, trust.
S5	dinner, play, set, return
S6	Enjoy, motivate, run, say.
S7	fall, feast, make, be, become, go, take, Drink, stop.
S8	Be, sing, excel, want, give.
S9	run, cake, eat, go, serve, unimpressed, reach, gives, visit, ask.
S10	discuss, beat, know, escape, bell, propose, suggested , tie, seemed ,ask, gather.

C] Extraction of Adjectives:

Adjectives are extracted from the stories. Adjective shows importance of a word naming an attribute of a noun, such as sweet, red, good , or technical . The Adjectives Extracted from the story. Extracted Adjectives as follows:

Table 6: Extracted Adjectives

Story Name	Extracted Adjectives
S1	Quick, good, hot, nearby, safe, last.
S2	Foolish, Remain, moral, adverse.
S3	Yet, certain, uncertain, Little, able, Small.
S4	Other, foolish, moral.
S5	Bad, moral, narrow-mounted, long- Necked, long, tough, good.
S6	Animal, interesting, little, tired, long, Powerful, strong, moral, valuable.
S7	Moral, hot, small, tired.
S8	Jealous, content, own, unique, Beautiful.
S9	stylish, moral, huge, high.
S10	easy, impossible, common, good, Young, cat, old, fine, moral.

D] Summary Generation*:* Depending upon extracted unique verb noun and Adjective, unique sentences has been generated which is act as the required sentence in summary generation of input text.

IV. EXPERIMENTAL RESULT

Experiment is performed by taking 10 input stories. such as S1,S2,S3,S4,S5,S6,S7,S8,S9,S10,Named Entity Recognition is applied as below. The current system produces sentences from unique verb, Noun and Adjectives_what the matter was. Selected sentences from 10 stories which reflect the important sentences necessary for the making summary.

A] S2: Once upon a time, a clever monkey lived on an apple tree. It was friends with a foolish crocodile that lived in the river. The monkey shared the fruits of the tree with the crocodile every day. The crocodile's wife learns about this friendship and asks the crocodile to bring the monkey's heart, which could be sweeter than the fruits of the tree. The couple invites the monkey for dinner and plan to eat his heart. The crocodile offers to take the monkey on its back, so that it can cross the river to reach home[20].

B] S5: There was once a fox, which was very friendly with a stork. It invited the stork to dinner one day and decided to play a prank ,So he set the table with a shallow dish, with little soup in it. The fox had a good meal, while the stork had a tough time drinking the tour with its long beak. The stork decided to return the kindness and invited the fox over for dinner and served soup in a long-necked, narrow-mouthed jar. This time the stork ate well, and the fox starved. Important factors Noun, verbs, Adjectives, location, and relation are covered in the sentences[20].

Most of Research work has focused on Extraction in Late 80s. The earliest instances of research on Automatic Text Summarization document proposed paradigms for extracting feature like Noun, Title, Association, Relation and Location from original text document . For present work small stories have been considered as input data. We have taken some stories of different sizes as input. We have to extract some text or important information from those stories, some word, sentences and phrases. We have to extract the feature of those stories like Title, Actors, Associations, Locations, Relations and Moral . The earliest instances of research on Automatic Text summarization document proposed paradigms for extracting Features of that document which is important.

Story name	Length of the story	Extracted unique sentence	Extracted actor	Ratio in percentage
S1	12	4	2	4/12=33%
S2	11	3	3	3/11=27%
S3	8	2	**2**	2/8=25%
S4	6	2	**2**	2/6=33%
S5	6	2	**2**	2/6=33%
S6	7	3	**2**	3/7=42%
S7	8	2	**2**	2/8=25%
S8	6	2	**2**	2/6=33%
S9	6	2	**2**	2/6=33%
S10	7	3	**2**	3/7=42%

Table 7 : Extracted unique sentences from the verb

V. CONCLUSION

The Vast growth in the rate of information due to internet has called for need of efficient text summarization systems. This paper has made for a very easy overview of information extraction from any text document in various steps. Nowadays, there are many ready-made software are available in market for information extraction process, like company name , employee details, location, area, Relationship, characters and activities etc . The information

DOI : 10.18231/2454-9150.2018.1301

extraction process is very important techniques for reader to get important points of some particular documents within fraction of seconds.

Acknowledgment

Authors would like to acknowledge and thanks to CSRI DST Major Project sanctioned No.SR/CSRI/71/2015(G), Computational and Psycholinguistic Research Lab Facility supporting to this work and Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India.

VII. References

[1].Raymond J. Mooney and Razvan Bunescu, june ,"Mining Knowledge from Text Using Information Extraction", Newsletter ACM SIGKDD Explorations Newsletter - Natural language processing and text mining, Volume 7 Issue 1. 2005.

[2]. “Raymond J. Mooney and Un Yong Nahm, September 2003,“Text Mining with Information Extraction.”, Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium, Bloemfontein, South Africa, Daelemans, W., du Plessis, T., Snyman, C. and Teck, L. (Eds.) pp.141-160, Van Schaik Pub., South Africa, 2005.

[3]. “C.P. Sumathi , T. Santhanam and G. Gayathri Devi, August 2012, “A SURVEY ON VARIOUS APPROACHES OF TEXT EXTRACTION IN IMAGES.” , International Journal of Computer Science & Engineering Survey (IJCSES) (IJCSES) Vol.3, No.4.

[4] Suneetha Manne, Zaheer Parvez Shaik Mohd. , Dr. S. Sameen Fatima, “Extraction Based Automatic Text Summarization System with HMM Tagger”, Proceedings of the International Conference on Information Systems Design and Intelligent Applications, 2012, Vol. 132, P.P 421-428.

[5] D. Sasirekha, E. Chandra , “Text Extraction from PDF document”, 2013, .Amrita International Conference of Women in Computing (AICWIC’13) Proceedings published by International Journal of Computer Applications® (IJCA).

[6] N.K. Gundu1, S.M. Jadhav, T.S. Kulkarni, A.S. Kumbhar, December 2014 “ Text Extraction from Image and Displaying its Related Information” , ISSN (Online): 2319-7064 , International Journal of Scientific and Research Publications, Volume 4, Issue 12, ISSN 2250-3153.

[7] Nimisha Dheer, Chetan Kumar ,2014, “ Automatic Text Summarization: A Detailed Study”, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064.

[8] Kumar Niraj, Srinathan Kannan and Varma Vasudeva, “A Knowledge Induced Graph-Theoretical Model for Extract and Abstract Single Document Summarization”, Computational Linguistics and Intelligent Text Processing - 14th International Conference, 2013.

[9] Mehdi Bazghandi, Ghamarnaz Tadayon Tabrizi and Majid Vafaei Jahan, “Extractive Summarization of Farsi Documents Based On PSO Clustering”, International Journal of Computer Science Issues, 2012, Vol. 9, Issue 4, No 3.

[10] Sivakumar A. P., Premchand P. and Govardhan A., “Query-Based Summarizer Based on Similarity of Sentences and Word Frequency”, International Journal of Data Mining & Knowledge Management Process (IJDKP), 2011, Vol.1, No.3.

[11].Ayush Agrawal , Utsav Gupta, “ Extraction based approach for text summarization using k-means clustering” International Journal of Scientific and Research Publications, Volume 4. November 2014.

[12] Anita .R. Kulkarni , Dr Mrs. S.S.Apte, Mar. - Apr. 2013, “ An automatic Text Summarization using feature terms for relevance measure”, IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 9, Issue 3.

[13]. “Vishal Gupta, Gurpreet Singh Lehal, AUGUST 2010 “A Survey of Text Summarization Extractive Techniques”, JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 2, NO. 3.

[14].“Sofien Lazreg, June 2012, “Using Information Extraction and Text Classification in an Effort to Support Systematic Literature Reviews” .

[15] “Deepali K. Gaikwad and C. Namrata Mahender, A Review Paper on Text Summarization” International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016.

[16]. Mehdi Bazghandi, Ghamarnaz Tadayon Tabrizi and Majid Vafaei Jahan, “Extractive Summarization of Farsi Documents Based On PSO Clustering”, International Journal of Computer Science Issues, 2012, Vol. 9, Issue 4, No 3.

[17]. Deshpande Anjali R., Lobo L. M. R. J., “Text Summarization using Clustering Technique”, International Journal of Engineering Trends and Technology (IJETT) , 2013, Vol. 4 Issue8.

[18]. Sivakumar A. P., Premchand P. and Govardhan A., “Query-Based Summarizer Based on Similarity of Sentences and Word Frequency”, International Journal of Data Mining & Knowledge Management Process (IJDKP), 2011, Vol.1, No.3.

[29]. Nikita Munot, Sharvari S. Govilkar, “Comparative Study of Text Summarization Methods”,

International Journal of Computer Applications (0975 – 8887) Volume 102– No.12, September 2014.

[20]. www.kidsworld.com moral stories of the kids world.

DOI : 10.18231/2454-9150.2018.1301