

Review Article

Testing Methodologies and Exploring Challenges and Issues in Text Summarisation

Darshna Patel^{1*} and Hitesh Chhinkaniwala²

¹Research Scholar, C.U. Shah University, Wadhwan, Surendranagar, Gujarat, India

²Associate Professor, Adani Institute of Infrastructure Engineering, Ahmedabad, Gujarat, India

(*Corresponding Author) Email-id: * darshnapatel@gmail.com; ² drhiteshrc1@gmail.com

ABSTRACT

Text summarisation is a text that is derived from one or more input text sources, which preserves an important portion of the information from the original text and significantly less than half of the original text. As the on-line information is increasing day-by-day on internet, users have to make more efforts to find the suitable information from online documents. This situation required research in the area of text summarisation within the Natural Language Processing (NLP) area. Early research started to summarise single document but as the amount of information increased, multi document summariser initiated, which gives summary from set of documents with the same topics or events. This survey studies mainly extractive, abstractive and hybrid approaches in text summarisation. This study intends to scrutinise some of the most relevant approaches for single- and multiple-document summarisations.

Keywords: Automatic text summarisation, Extractive summarisation, Abstractive summarisation, Natural language processing, Evaluation measures, Statistic approaches, Linguistics approaches

1. INTRODUCTION

Text summarisation is the abridge version of one or more documents into a summarised version by extracting almost all the silent concepts by applying different text mining techniques. As an explosive growth of textual information available on online documents, the users have to put extra efforts to find out interested information. This requires an automatic tool or technique, when user entered a text in to the computer, a non-redundant text from original text is returned called text summarisation. Systematically text summarisation can be distinguished among following types of summaries: extractive/abstractive, single-/multi-document, generic/query-based and mono-lingual/multi-lingual/cross-lingual. Most existing summarisers work in an extractive fashion, extracts the important

sentences of the input documents that are believed to be more salient. Abstraction approach is to derived summary by paraphrasing sections of the original document. This approach is more powerful and generates close result to human generated result but it required deeper understanding of the input text and so limited to small domain. Query-based summaries are relying on the user specification while generic summaries attempt to identify the important information in document without the context of a query.

2. TEXT SUMMARISATION

The past literatures that use the various summarisation approaches are basically of extractive and abstractive approaches. Some of researchers also use some mathematical formulation techniques like algebraic

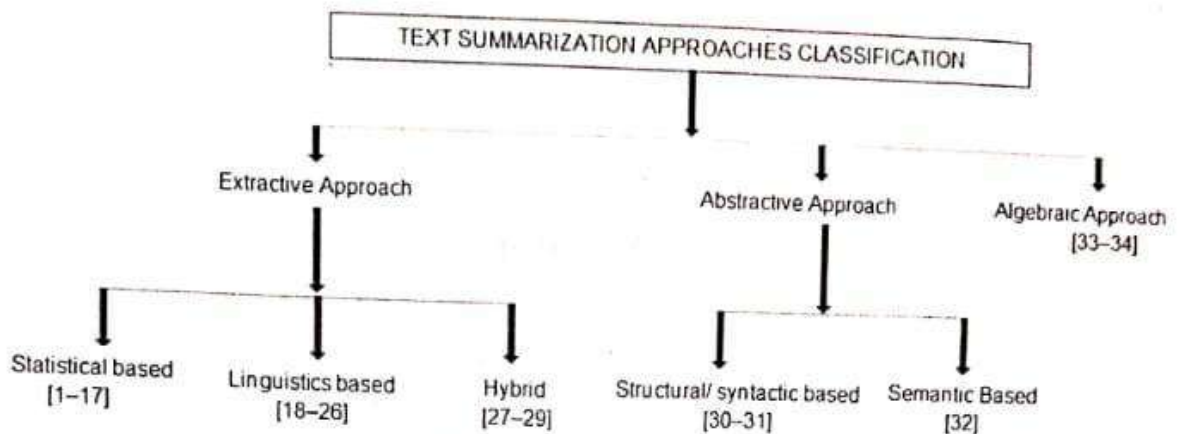


Figure 1: Classification of text summarisation approaches

reduction approach for the text summarisation. The most widely used methods for summarisation is based on statistical and linguistics methods which score and rank the sentences in the document and generates the summary. Steps to text summarisation involve elimination of redundancy, identification of importance sentences, generating coherent summary and evaluating automatic generated summary. Various techniques for text summarisation are shown in Figure 1.

3. EXTRACTIVE SUMMARISATION

This approach aims to pick out most relevant and significant sentences in the document with minimum redundancy. The key ideas in this approach are performing the study analysis on human generated summary and determine the characteristics expected in automatic generated summary.

There are mainly two approaches to implement extractive summary.

3.1 Statistic-Based Approaches

Traditional methods rely on the statistical methods to find essential concepts from the documents. Statistical methods calculate the score of the sentences based on statistical features of the sentences. Various sentence features used for sentences scoring are classified into word- and sentences-based and graph level scoring as shown in Figure 2^[1,2].

Effectiveness of different sentence scoring method depends on the kind of input sources, length of text and kind of language used and their structure^[1]. To advocates this thesis Ferreira *et al.*^[1] suggested context-dependent summarisation system which brings evidence of this thesis by combining different sentence scoring methods

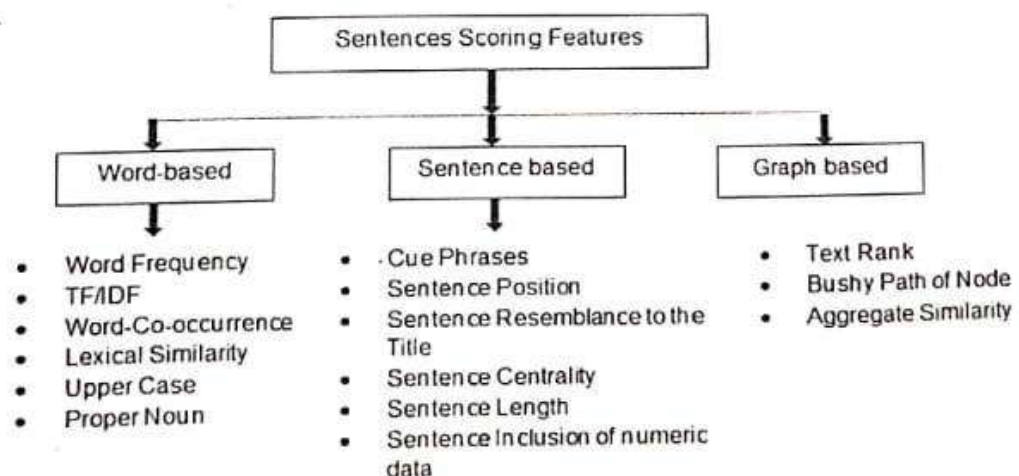


Figure 2: Classification of sentence scoring features

used in different context which can be further improved by developing system that automatically analyses the features of input text and automatically discovers best combination of sentence scoring. In [2], authors also analysed that sentence centrality, sentence location, named entity and proper noun are best option to create feature set and suggested that summary can be improved by adding semantic features and weighting feature by genetic algorithm. Mendoza *et al.* [3] developed algorithm MA-Single DocSum for monolingual document of any genre based on genetic operator algorithm which weighting individual statistical features of each sentence. Chandra *et al.* [4] developed *K*-mixture probabilistic model to establish term weight in statistical sense and identify the term relationships to derive the semantic relationship significance of nouns that manifests sentence semantics. Limitation of this model extracts only noun as features.

Various techniques have been explored that consist of extracting term or keyword from single- or multi-documents, ranking and scoring sentences and generating summary with higher score sentences with minimum overlapping [5-12]. Mohamed Abdel Fattah [5] constructed supervised text summariser model based on maximum entropy model, support vector machine and a naïve Bayes classifier which extracts features in training mode. Mei and Chen [6] introduced SumCR subtopic-based approach for multi-document which uses clustering to discover subtopics and two features 'Exemplar' and 'Position' incorporated for sentence scoring. The same approaches can also be used to aim to multilingual summarisation found in PLIS [7] which combines techniques for language identification, content translation and summarisation. Text summarisation task consists of various subtask like identification of relevant sentences and so on and each of that tasks directly affects the performance of summarisation, with this consideration. Babar and Patil [8] proposed a new method which integrates the Fuzzy logic to extract the semantic relation between concepts of text for single-document summarisation. In [9], the authors have investigated use of genetic algorithm (GA), Mathematical Regression (MR), Feed forward neural network (FFNN) and probabilistic neural network (PNN) to address the problem of improving content selection in text summarisation. In [10], the authors have proposed Arabic text summarisation system which uses clustering and have adapted the discriminant analysis method to score the term with minimum redundancy and maximum relevance. One unsupervised and domain-

independent approach is found in [11], which uses graph-based ranking model for text summarisation. Ko and Seo [12] have proposed an effective hybrid method to extract salient sentences using contextual information with statistical approach to solve the features sparseness problem for both single- and multi-document summarisation.

Another work found in literature that aims to multilingual summarisation in PLIS [13] proposed a language-independent summarisation that combines techniques for language identification, content translation and summarisation. Ghalehtaki *et al.* [14] proposed a combination method of Particle Swarm Optimisation, Cellular Learning Automata to construct their model, which could recognise sentence richness and can build summary which is close to human generated summary. Fang *et al.* [15] proposed topic aspect-oriented summarisation approach for feature extraction with the consideration that summary from multi-document of one topic can describe various aspects of one given topic. Reference [16] presents technique for generating single web document summary based on frequent terms and words in the corpus. The same approach can also be applied to multi-document summarisation. Further improvement can be done by applying concept of semantic similarity to extract the words. Yirdaw and Ejigu [17] proposed concept-based approach for Amharic text single-document summarisation. They show that a principled use of term by concept matrix resulted from Probabilistic Latent Semantic Analysis (PLSA) model can help to produce summary that captures the main topics of document.

The limitation of statistical approach is that it does not consider the semantic relationships among sentences which may cause summary inconsistent and incoherent. To remove the limitation of these statistical approaches, many researchers focused on applying semantic analysis of text.

3.2 Linguistics-Based Approaches

Linguistics-based approach looks into the term semantics, it identifies the semantic relationship between terms, words, phrases and structural information in the documents using linguistics resources like Part of Speech (POS), tagging, grammar analysis and thesaurus usage [4,33].

The study of Term-pair co-occurrence is a very successful application in information retrieval. The main idea of this study is that if the co-occurrence rates of two terms

are larger than those, two terms are semantically related. Based on this idea Liu *et al.* [18] proposed multi-document-based summarisation to utilise latent semantic relation between terms in document and established topic-oriented database of word co-occurrence. Xu [19] proposed a method for multi-topic text partitioning based on semantic distance between paragraphs and lastly extracts key sentences from different topics.

In traditional clustering, algorithm does not consider the semantic relationship among words. To solve this issue, semantic information from WordNet used to improve the quality of clustering. Zhang and Li [20] proposed sentence clustering approach for text summarisation by identifying semantic relations among sentences. Pal and Saha [21] performed summarisation task by performing semantic analysis of the text. In the approach, simplified Lesk algorithm is used for extracting relevant sentences, and WordNet is used as semantic dictionary. But still problems such as synonym, polysemy, high dimensionality exist. Tingting wei *et al.* [22] proposed methodology for clustering using WordNet and Lexical chain where Lexical chains are used to extract core semantic features that express the topic of document. Some limitations of this approach are that important words which are not included in WordNET lexical will not be considered as concept for similarity evaluation. Most existing graph-based ranking algorithm ignores the semantic information of sentences which can provide some clues for sentence ranking in summarisation. Su and Wan [23] proposed novel graph-based sentence ranking algorithm SRRank to incorporate the semantic role information in to graph-based ranking algorithm. This method can be improved by using Head-driven Phrase Structure Grammar (HPSG) parser to derive more semantic information. To grabbing the coherent sentences, Sankar and Sobha [24] proposed text ranking approach based on graph ranking model which involves two steps of operation for summary. First, they find the coherent chunks in document and in the second step rank the sentences in individual chunks. Another interesting and important methodology of information retrieval from webpages of textual information is found in [25]. This multi-lingual approach uses natural language processing (NLP), semantics, statistics and fusion techniques to obtain informative summary from web pages. Alguliev *et al.* [26] have innovated multi-document summarisation

model with ability to remove redundant information as a discrete optimisation problem. They have created adaptive differential algorithm to solve the optimisation problem.

3.3 Hybrid Approaches

In the above sections, we have discussed various methods based on statistical and linguistics features of sentences. To get the better summary of single- and multi-document, both statistical and linguistics features of sentences are important for sentences scoring and ranking. With this thought PadmaLahri *et al.* [27] proposed a new method for summarisation which uses statistical features as well as linguistics features of sentences. The effectiveness of summary can be achieved by adding some more features like sentence-to-sentence cohesion, biased words and so on.

The main issue in multi-document summarisation system is redundancy and information diversity. To deal with these issues, Ferreira *et al.* [28] proposed new graph-based clustering algorithm which can represent document using statistical similarity, semantic similarity, co-reference and discourse relations. The limitation of this proposed approach is sentence ordering because it finds sentences in groups of different topics. This system can be made as multi-lingual system by adding language detection component. Dalal and Zaveri [29] also proposed combinational methods of heuristics to generate single-document summary.

4. ABSTRACTIVE SUMMARISATION

Still two main problems exist in extractive summarisation. First, textual coherence and redundant phrases exists in summary. Automatic summarisation can solve these problems by carrying NLP. Abstractive summarisation needs deeper understanding of text and is able to generate new sentences from text which improves text summary.

Abstractive summarisation techniques are categorised in two approaches, structured-/syntactic- and semantic-based approach as shown in Figure 3.

4.1 Syntactic/Structural-Based Approaches

Recent approaches in abstractive summarisation use word graph for representing documents. Le *et al.* proposed [30] abstractive approach for summarisation works in two steps: sentence reduction which is based on discourse rules

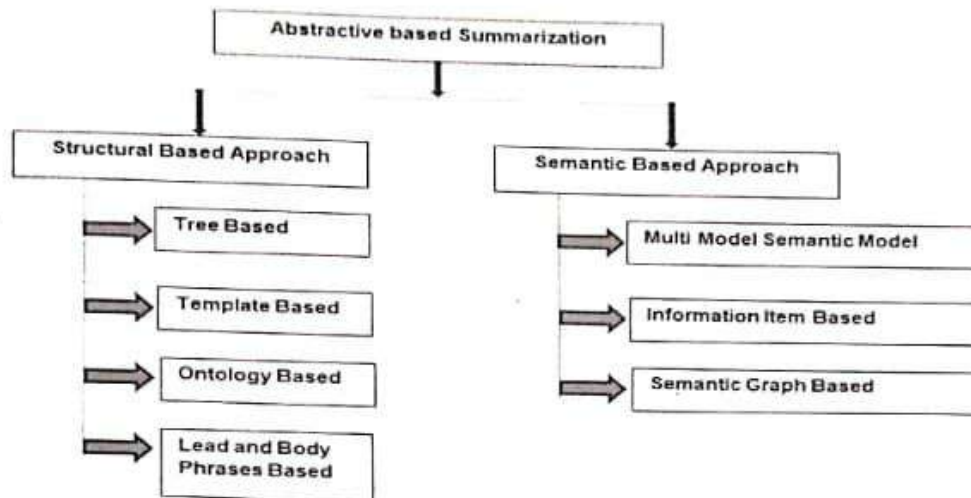


Figure 3: Classification of abstractive summarisation

and syntactic constraints, sentence combination which is based on word graph to represent relation among words, phrases and sentences. Many researchers have made effort to use ontology or knowledge base to improve the process of summarisation. In [31], Texminer works on the socio-cognitive approach, the structural discourse models and the ontology-based summarisation.

4.2 Semantic-Based Approaches

In semantic-based method, semantic representation of document(s) is used to feed into Natural Language Generation (NLG) system. In literature, abstractive summarisation is based on Rich Semantic Graph for source document. Rich Semantic Graph is an ontology-based representation and limitation of this approach is that it depends on the human expert created ontology and rules which is time-consuming effort. Aims to treat this limitation [32] proposed a framework that used semantic rule labelling for semantic representation of text to generate good summary.

5. ALGEBRAIC REDUCTION APPROACHES

In this section, we discuss the recent development in the mathematical foundation approach to identify and extract the most important information in document. Information retrieval using mathematical foundation approach started with Boolean model, probabilistic model and vector space model [29]. Mainly we focus on algebraic reduction methods.

Several algebraic reduction methods are used to recognise and extract the semantically important text in document. Reference [33] focused on the application of Single Value Decomposition (SVD) and Non-negative Matrix Factorization (NMF) based approaches on automatic summarisation with their merits and demerits. Vector Space Model (VSM) transforms the content of the text corpus in to term-document matrix. Two main issues with VSM are its incapability in capturing semantic information and linguistic noise. This limitation can be overcome by latent semantic analysis which can be used for detecting underlying semantic relationship within text. The major drawback of SVD is the possibility of negative value occurrence. This problem can be solved by NMF. NMF can provide better conceptual representation of data than SVD. Semantic variable vector achieved by NMF are sparser than SVD. NMF has its convergence issue which is to extract correct feature. Ahn *et al.* [34] also used simplicial NMF method in their unsupervised generic document summarisation model which inherits advantages of NMF and furthermore used Maximal Marginal Relevance (MMR) matrix to remove the redundancy from summary.

6. SUMMARY OF LITERATURE SURVEY

Summary of different text summarisation approaches, summary type with possible improvement used in previous researches are shown in Table 1.

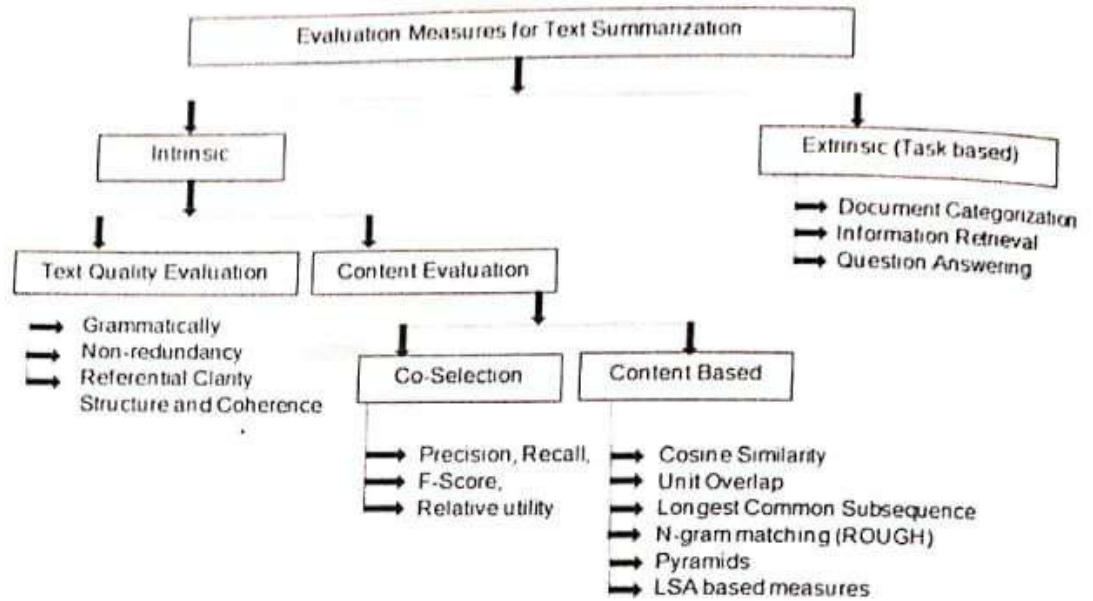


Figure 4: Different evaluation measures for text summarisation

Table 1: Summary of different text summarisation approaches used in previous researches

Ref. No.	Authors (Year)	Approaches Used	Summarisation Method and Summary Type	Findings and Possible Improvement
[1]	Ferreira R, Freitas F, de Souza Cabral L, Dueire Lins R, Lima R, Franca G <i>et al.</i> (2014)	Combine different sentences scoring as per context	Extractive Statistical Unsupervised Single document	If kind of input source changes, result may affect To develop a system which automatically analyses the features of input sentences and automatically combines the best sentences scoring methods
[2]	Meena YK, Gopalani D (2014)	Sentence score based on sentence features	Extraction Statistics Single document	Authors have used features simply without any weight associated Features can be included with weighting system and genetic algorithms can be used to get proper weight We can add some semantic features and WordNET to current version of summariser to improve performance
[3]	Mendoza M, Bonilla S, Noguera C, Cobos C, León E (2014)	Weighting statistical features of sentences	Extraction Statistical Single document	Work for Single document and monolingual summarisation 1. Plan to make proposal for query based single document summarisation using proposed approach. 2. Evaluate proposed memetic algo. With other dataset 3. To involve study of other schemes of selection, crossover, mutation, replacement, and local search in order to obtain a new configuration for memetic algorithm
[4]	Chandra M, Gupta V, Paul SK (2011)	Extract essential concepts of document, text categorisation	Extraction Statistical Single document	Only extract noun as features Better summarisation approach form better summary

Table 1: continued

Ref. No.	Authors (Year)	Approaches Used	Summarisation Method and Summary Type	Findings and Possible Improvement
[5]	Fattah MA (2014)	Content selection/feature extraction	Extraction Statistical Multi document	We can add graph-based features to improve the quality of summary
[6]	Mei J-P, Chen L (2012)	Sentence extraction, capture subtopic	Extraction Statistical Multi-document	Used only keyword-based similarity measure so not incorporate query-based Summarisation Query-based Summarisation can be improved by using advanced query processing techniques and sophisticated sentences similarity measures
[7]	Cabral LD, Lins RD, Mello RF, Freitas F, Ávila B, Sinske S, Riss M (2014)	Statistical Sentence scoring	Extraction Statistical Multi lingual	Work on textual information, Language focused are from European Union To encompass other language and summarisation method for better result
[8]	Babar SA, Patil PD (2015)	Feature extraction	Extraction Statistical Single document	Focus of this paper is narrow, limited to single document Extend proposed method for multi document summarisation with large dataset and domain specific data
[9]	Fattah MA, Ren F (2009)	Feature extraction	Extraction Statistical Single document	Single document summarisation Extend proposed method for multi document summarisation by addressing some anti-redundancy methods
[10]	Oufaïda H, Nouali O, Blache P (2014)	Scoring terms	Extraction Statistical Single and multi-document	Single document summarisation leads to coherent problem. Multi document Sentence reordering is issue and not considered in present paper Intend to conduct experiments in other domain and investigate the cross lingual task
[11]	Hamid F, Tarau P (2014)	To extract keyword/ Sentences	Extraction Statistical Single document	Plan to combine some rule-based algorithm which will help resolute anaphora between nouns and pronouns in the following sentences
[12]	Ko Y, Seo J (2008)	To extract salient sentences	Extraction Statistical Single and multi-document	Statistical methods do not use semantic features of sentences, Extraction method may cause cohesion problem Use of linguistics methods may improve summary
[13]	Nandhini K, Balasundaram SR (2013)	Extract important and readable sentences	Extraction Statistical Single document	Lack of cohesion is major drawback Improving Cohesion and extracting individualised summary
[14]	Ghalehtaki RA, Khotanlou H, Esmailpour M (2014)	Sentence similarity, sentence weight, score to sentences	Extraction Statistical Single document	Worked only for single document Same approach can be extended for multi document
[15]	Fang H, Lu W, Wu F, Zhang Y, Shang X, Shao J, Zhuang Y (2015)	Sentence feature extraction and feature group construction	Extraction Statistics Multi document	Extract three kinds of sentence features TAOS outperforms when stop words are removed

Table 1: continued

Ref. No.	Authors (Year)	Approaches Used	Summarisation Method and Summary Type	Findings and Possible Improvement
[16]	Bagalkotkar A, Kandelwal A, Pandey S, Kamath SS (2013)	Frequent terms based	Extraction Statistical Single document	Concept of semantic similarity is not applied to frequent terms Same concept can be applied to multi document texts Semantic similarity concept can be applied so sentences with frequent terms and semantically similar terms can be extracted
[17]	Yirdaw ED, Ejigu D (2012)	Concept-based topic modelling with PLSA	Extraction Statistical Single document	Amharic Text does not contain Linguistics features so not considered that features Do not used Term weighting Techniques for PLSA Applying Proposed work for Multi document, query focused summarisation Find out Term Weighting Techniques for PLSA Establishment of Large Scale Data for Evaluation
[18]	Liu N, He Y, Chen Q, Peng M, Fang W (2013)	Frequency of term-pair co-occurrence	Extraction Linguistics Methods Multi document	Effective with the text written in traditional structure Text with new Structure required further study
[19]	Xu X (2009)	Semantic distance of paragraph	Extraction Linguistics Methods Single document	Readability and coherence of context is poor Use of NLP to improve quality of summary specially readability
[20]	Zhang PY, Li CH (2009)	Semantic distance of sentences Clustering	Extraction Linguistics Methods Single document	Considered only three features of sentences Word Form, word Order Form and Semantic features More features of sentences to rank the sentences for better summary
[21]	Pal AR, Saha D (2014)	Sentences weight	Extraction Linguistics Based Unsupervised Single document	Hand rules are not considered Hybridisation of proposed approach with some hand tagged rule can improve the TS quality
[22]	Wei T, Lu Y, Chang H, Zhou Q, Bao X (2014)	Text clustering Semantic approach using WordNet and lexical chains	Extraction Linguistics Single document	Important words which are not included in WordNET lexicon will not considered as concepts for similarity evaluation Perform this method on larger knowledge base such as Wikipedia
[23]	Yan S, Wan X (2014)	Topic focused ranking sentences	Extraction Linguistics Multi document	When weight of outer is too large (>0.5) performance becomes poor Use of other semantic parser like HPSG parser to derive more semantic role information Integrate SRRank-cluster and SRRank span for better result
[24]	Sankar K, Sobha L (2009)	Coherence chunker and hash based approach to graph-based unsupervised ranking algorithm	Extraction Linguistics Single document	Work only for single document summarisation Work for multi document summarisation
[25]	Escudero S, Garrido AL, Ilarri S (2014)	Web fusion techniques, NLP, semantics, statistics	Extraction Linguistics Multilingual	Work only on Text-based information To make system interface more user friendly and interactively

Table 1: continued

Ref. No.	Authors (Year)	Approaches Used	Summarisation Method and Summary Type	Findings and Possible Improvement
[26]	Alguliev RM, Aliguliyev RM, Mehdiyev CA (2011)	Model as optimisation problem	Extraction Centroid based Multi document	Vector Space Model is used for sentence similarity which has high dimensionality of feature space which is drawback of VSM To solve optimisation problem with diff algorithm, Particle swarm optimisation, ant algorithm and differential evolution algo with diff mutation strategies
[27]	Padma Lahari E, Siva Kumar DVN Prasad S (2014)	Sentence score	Extraction Statistical and linguistics Single document	Used limited Sentence Features Quality of TS can be achieved by adding some advanced features such as sentence to sentence cohesion, biased words etc.
[28]	Ferreira R, de Souza Cabral L, Freitas F, Lins RD, de França Silva G, Simske SJ <i>et al.</i> (2014)	Sentence clustering based on graph model	Extraction Statistical and linguistics methods Unsupervised Multi document	Sentence ordering because sentences are selected from different group of topic Sentences can be order by its relative position in original document Adding Language detection component, can make this system multilingual, multi document summarisation system
[29]	Dalal MK, Zaveri MA (2011)	Heuristics based	Extraction Statistical and linguistics Single document	Limited to single document summarisation Practical application like automatic grading, required study of human cognitive process and machine learning methods
[30]	Le HT, Le TM (2013)	Sentence reduction and sentence combination	Abstractive approach, Syntactic based Single document	Input of system is extractive summary (I) propose methods to improve the meaning completeness of sentences generated in the sentence reduction phrase; (ii) propose methods to further compress sentences; and (iii) investigating strategies to efficiently combine sentences in the summary
[31]	Hipola P, A. Senso J, Leiva-Mederos A, Domínguez-Velasco S (2014)	Ontology based	Abstraction Syntactic based Multi document	inefficient in different knowledge domain Make system that can be used in different domains
[32]	Khan A, Salim N, Kumar YJ (2015)	Based on semantic role labelling	Abstractive approach, Semantic based Multi document	Precision value of proposed framework drops as compare to other system Integrate graph with SRL to build a semantic graph for multi-document abstractive summarisation
[33]	Batcha NK, Zaki AM (2010)	High to low dimension data SVD, NMF	Mathematical foundation Single document	Drawback of SVD is the possibility of negative value occurrence NMF has convergence issue to extract correct features Mixing anaphora resolving with proper parameter values of NMF like ranking, initialisation, iterations and tolerance can be used to achieve favourable result
[34]	Anh NK, Toi NK, Linh NV (2014)	Identify major Topic from sentences with minimum redundancy	Mathematical foundation Single document	Work only for single document Proposed method can be extended for multi document summarisation

7. EVALUATION MEASURES OF TEXT SUMMARISATION

Evaluation measures determine the quality of summary with respect to different parameters like structure, length, coherence and content. Basically two parameters are very important in text summarisation: Compression Ratio and Retention Ratio which determine how much shorter the summary and how much of central information preserves, respectively. A broad classification of evaluation techniques is shown in Figure 4.

7.1 Text Quality Measures

Grammatically: The text should not contain non-textual items.

Non-redundancy: The text should not contain redundant information.

Reference clarity: The nouns and pronouns should be clearly referred to in the summary.

Coherence and structure: The summary should have good structure and the sentences should be coherent.

This cannot be done automatically. The annotators mostly assign marks (i.e. from A – very good to E – very poor at DUC 2005) to each summary.

7.2 Co-Selection Measures

Precision: Precision (P) is the number of sentences occurring in both system and ideal summaries divided by the number of sentences in the system summary.

Recall: Recall (R) is the number of sentences occurring in both system and ideal summaries divided by the number of sentences in the ideal summary.

F-score: F-score is a composite measure that combines precision and recall.

$$F = \frac{2 \cdot P \cdot R}{P + R} \quad (1)$$

Below is a more complex formula for measuring the F-score

$$F = \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R} \quad (2)$$

where β is a weighting factor that favours precision when $\beta > 1$ and favours recall when $\beta < 1$.

Relative utility: The model summary represents all sentences of the input document with confidence values for their inclusion in the summary

$$RU = \frac{\sum_{j=1}^n \delta_j \sum_{i=1}^N u_{ij}}{\sum_{j=1}^n \in_j \sum_{i=1}^N u_{ij}} \quad (3)$$

where u_{ij} is a utility score of sentence j from annotator i , δ_j is 1 for the top ε sentences according to the sum of utility scores from all judges, otherwise its value is 0, and $\in_j$ is equal to 1 for the top ε sentences extracted by the system, otherwise its value is 0.

7.3 Content-Based Measures

Cosine similarity:

$$\cos(X, Y) = \frac{\sum_i x_i \cdot y_i}{\sqrt{\sum_i (x_i)^2} \sqrt{\sum_i (y_i)^2}} \quad (4)$$

where X and Y are representations of a system summary and its reference document based on the vector space model.

Unit overlap:

$$\text{Overlap}(X, Y) = \frac{\|X \cap Y\|}{\|X\| + \|Y\| - \|X \cap Y\|} \quad (5)$$

where X and Y are representations based on sets of words or lemmas. $\|X\|$ is the size of set X

Longest common subsequence:

$$\text{lcs}(X, Y) = \frac{\text{length}(X) + \text{length}(Y) - \text{edit } d_i(X, Y)}{2} \quad (6)$$

where X and Y are representations based on sequences of words or lemmas, $\text{lcs}(X, Y)$ is the length of the longest common subsequence between X and Y , $\text{length}(X)$ is the length of the string X , and $\text{edit}_{di}(X, Y)$ is the edit distance of X and Y

n-Gram matching (ROUGH) =

$$\text{ROGUE}_n = \frac{\sum_{c \in \text{RSS}} \sum_{\text{gram}_n \in C^{\text{Count}} \text{match}(\text{gram}_n)} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{c \in \text{RSS}} \sum_{\text{gram}_n \in C^{\text{Count}}(\text{gram}_n)} \text{Count}(\text{gram}_n)} \quad (7)$$

Where $\text{Count}_{\text{match}}(\text{gram}_n)$ is the maximum number of n -grams co-occurring in a candidate summary, and a reference summary and $\text{Count}(\text{gram}_n)$ is the number of n -grams in the reference summary.

Pyramids: It is semi-automatic evaluation method and its basic idea is to identify summarisation content units (SCUs) that are used for comparison of information in summaries.

7.4 Extrinsic (Task-Based Measures)

Document categorisation: This evaluation seeks to determine whether the generic summary is effective in capturing whatever information in the document is needed to correctly categorise the document.

Information retrieval: If a summary captures the main points of a document, then an IR machine indexed on a set of such summaries (instead of a set of the full documents) should produce (almost) as good a result.

Question answering: The purpose of this measure is to test the readers' understanding of the summary and its ability to convey key facts of the source article.

8. RESEARCH FINDINGS

In extractive summarisation approach, readability and coherence is poor and becomes more severe in multi-document summarisation where information is drawn from different sources. Post-processing is required to address this issue.

Most of the methods are dependent on sentences features and hand tagged rules for sentences scoring so kind of input source changes can affect the result. It requires developing a system which automatically analyses the features of input sentences and automatically combines the best suitable sentences scoring methods depending on input source.

Most of extractive summarisation approaches are effective only with the text written in traditional structure. Articles which contain figures and tables are not considered in summarisation process. By applying semantic similarity to sentence features and sentence similarity, sentence-to-sentence cohesion, biased word, graph-based features summary can improve.

- Summary can be more improved by adding weighting features using Genetic Algorithm (GA).
- Query-based text summarisation can be improved by adding query processing techniques and sophisticated sentences similarity measure.
- Scale of data corpus is not considered.
- By adding language detection tool, multilingual system can be developed.
- Making summarisation system that can be are used friendly and interactive more interactively.
- There is no generalised framework that humans can use for abstractive summarisation.

9. CONCLUSION

Explosion of textual information on World Wide Web has called for a need to develop text summarisation system. Research on text summarisation has drifted from technical articles to news articles, Emails and blogs and single- to multi-documents and multi-lingual documents. To create summary, both abstractive and extractive approaches are attempted where abstractive approach required deeper knowledge for language generation and challenging area because of complexity of NLP. In this paper, we only mention the abstractive methods and focus on the extractive summarisation methods. An extractive-based method selects the importance of sentences from the original text based on the statistical and linguistic features of sentences.

REFERENCES

- [1] Ferreira R, Freitas F, de Souza Cabral L, Dueire Lins R, Lima R, Franca G *et al.* A context based text summarization system. Document analysis systems (DAS), 2014 11th IAPR international workshop on. IEEE; 2014.
- [2] Meena YK, Gopalani D. Analysis of sentence scoring methods for extractive automatic text summarization. Proceedings of the 2014 international conference on information and communication technology for competitive strategies. ACM; 2014.
- [3] Mendoza M, Bonilla S, Noguera C, Cobos C, León E. Extractive single-document summarization based on genetic operators and guided local search. Expert Sys Appl 2014;41(9):4158–69.
- [4] Chandra M, Gupta V, Paul SK. A statistical approach for automatic text summarization by extraction. Communication systems and network technologies (CSNT), 2011 international conference on. IEEE; 2011.
- [5] Fattah MA. A hybrid machine learning model for multi-document summarization. Appl Intell 2014;40(4):592–600.
- [6] Mei J-P, Chen L. SumCR: a new subtopic-based extractive approach for text summarization. Knowl Inf Syst 2012;31(3):527–45.
- [7] Cabral LD, Lins RD, Mello RF, Freitas F, Ávila B, Simske S, Riss M. A platform for language independent summarization. Proceedings of the 2014 ACM symposium on document engineering. ACM; 2014.
- [8] Babar SA, Patil PD. Improving performance of text summarization. Procedia Comput Sci 2015;46:354–63.
- [9] Fattah MA, Ren F. GA, MR, FFNN, PNN and GMM based models for automatic text summarization. Comput Speech Lang 2009;23(1):126–44.
- [10] Oufaida H, Nouali O, Blache P. Minimum redundancy and maximum relevance for single and multi-document Arabic

- text summarization. *J King Saud Univ Comput Inf Sci* 2014;26(4):450–61.
- [11] Hamid F, Tarau P. Text summarization as an assistive technology. *Proceedings of the 7th international conference on pervasive technologies related to assistive environments*. ACM; 2014.
- [12] Ko Y, Seo J. An effective sentence-extraction technique using contextual information and statistical approaches for text summarization. *Pattern Recognit Lett* 2008;29(9):1366–71.
- [13] Nandhini K, Balasundaram SR. Improving readability through extractive summarization for learners with reading difficulties. *Egypt Inform J* 2013;14(3):195–204.
- [14] Ghalehtaki RA, Khotanlou H, Esmailpour M. A combinational method of fuzzy, particle swarm optimization and cellular learning automata for text summarization. *Intelligent systems (ICIS)*, 2014 Iranian conference on. 4–6 Feb., 2014. p. 1–6.
- [15] Fang H, Lu W, Wu F, Zhang Y, Shang X, Shao J, Zhuang Y. Topic aspect-oriented summarization via group selection. *Neurocomputing* 2015;149:1613–9.
- [16] Bagalkotkar A, Kandelwal A, Pandey S, Kamath SS. A Novel Technique for Efficient Text Document Summarization as a Service. In *Advances in Computing and Communications (ICACC)*, 2013 Third International Conference on 2013 Aug 29 (pp. 50–53). IEEE.
- [17] Yirdaw ED, Ejigu D. Topic-based Amharic text summarization with probabilistic latent semantic analysis. *Proceedings of the international conference on management of emergent digital EcoSystems*. ACM; 2012.
- [18] Liu N, He Y, Chen Q, Peng M, Fang W. Multi-document Biased Summarization based on topic-oriented characteristic database of term-pair co-occurrence. *Information science and technology (ICIST)*, 2013 international conference on. IEEE; 2013.
- [19] Xu X. Research on automatic summarization system based on topic partition. *Web information systems and mining, 2009. WISM 2009. International conference on*. IEEE; 2009.
- [20] Zhang P-Y, Li C-H. Automatic text summarization based on sentences clustering and extraction. *Computer science and information technology, 2009. ICCSIT 2009. 2nd IEEE international conference on*. IEEE; 2009.
- [21] Pal AR, Saha D. An approach to automatic text summarization using WordNet. *Advance computing conference (IACC)*, 2014 IEEE international. IEEE; 2014.
- [22] Wei T, Lu Y, Chang H, Zhou Q, Bao X. A semantic approach for text clustering using WordNet and lexical chains. *Expert Syst Appl* 2015;42(4):2264–75.
- [23] Yan S, Wan X. SRRank: leveraging semantic roles for extractive multi-document summarization. *IEEE/ACM Trans Audio Speech Lang Process* 2014;22(12):2048–58.
- [24] Sankar K, Sobha L. An approach to text summarization. *Proceedings of the third international workshop on cross lingual information access: addressing the information need of multilingual societies*. Association for Computational Linguistics; 2009.
- [25] Escudero S, Garrido AL, Harri S. Obtaining knowledge from the Web using fusion and summarization techniques. *Information fusion (FUSION)*, 2014 17th international conference on. IEEE; 2014.
- [26] Alguliev RM, Aliguliyev RM, Mehdiyev CA. Sentence selection for generic document summarization using an adaptive differential evolution algorithm. *Swarm Evol Comput* 2011;1(4):213–22.
- [27] PadmaLahari E, Siva Kumar DVN Prasad S. Automatic text summarization with statistical and linguistic features using successive thresholds. *Advanced communication control and computing technologies (ICACCCT)*, 2014 international conference on. IEEE; 2014.
- [28] Ferreira R, de Souza Cabral L, Freitas F, Lins RD, de França Silva G, Simske SJ *et al*. A multi-document summarization system based on statistics and linguistic treatment. *Expert Syst Appl* 2014;41(13):5780–7.
- [29] Dalal MK, Zaveri MA. Heuristics based automatic text summarization of unstructured text. *Proceedings of the international conference & workshop on emerging trends in technology*. ACM; 2011.
- [30] Le HT, Le TM. An approach to abstractive text summarization. *Soft computing and pattern recognition (SoCPar)*, 2013 international conference of. IEEE; 2013.
- [31] Hipola P, A. Senso J, Leiva-Mederos A, Domínguez-Velasco S. Ontology-based text summarization. *The case of Texminer. Library Hi Tech* 2014;32(2):229–48.
- [32] Khan A, Salim N, Kumar YJ. A framework for multi-document abstractive summarization based on semantic role labelling. *Appl Soft Comput* 2015;30:737–47.
- [33] Batcha NK, Zaki AM. Algebraic reduction in automatic text summarization – the state of the art. *Computer and communication engineering (ICCCE)*, 2010 international conference on. IEEE; 2010.
- [34] Anh NK, Toi NK, Linh NV. An interpretable method for text summarization based on simplicial non-negative matrix factorization. *Proceedings of the fifth symposium on information and communication technology*. ACM; 2014.