International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 23 (2015) pp 43501-43508
© Research India Publications. http://www.ripublication.com

Topic Ontology Assisted Multi-document Summary Generation

K.Yogeswara Rao
Department of Computer Science and Engineering,
GITAM University,
Visakhapatnam,
Andhrapradesh.
yogiindusisu@gmail.com

P.V.Nageswara Rao
Department of Computer Science and Engineering,
GITAM University,
Visakhapatnam,
Andhrapradesh.
nagesh@gitam.in

Abstract

An effective Multiple Document Summarization (MDS) system is a sound method to provide concise and comprehensive information in a short-form. The conventional summarization techniques exploit the machine learning techniques based sentence extraction and sentence position hypothesis to summarize the huge document collection under the same topic. However, these techniques may lead the redundant and less informative sentences in the summary due to the lack of semantic analysis. To tackle this constraint, this paper attempts to exploit the ontology and word position hypothesis. This paper proposes multi-document SummarY generatioN with the tOPic ontology asSIStance (SYNOPSIS) approach. The core aim of this approach is to balance the objective function that refers to improve the coherence, and salience of the summary and diminish the redundancy of the sentences. To achieve this aim, the SYNOPSIS model investigates, the two phases, namely optimal sentence ranking and sentence selection in MDS system. Initially, the SYNOPSIS uses Yago ontology to identify the context of the keyword semantically and employs the word position hypothesis to discover the importance of the sentence by ranking the document sentences. To further reduce the document content, the SYNOPSIS approach focuses on shortening the sentence length by applying the structure analysis of the original sentences. Finally, it selects the key sentences based on the most relevant information on sentence rank and constructs the summary based on the satisfaction of the objective function. The experimental results demonstrate that the SYNOPSIS approach achieves better performance than the conventional summarization method.

Keywords: Ontology, summary, word position, document summarization, coherence, salience, and redundancy

1. Introduction

Due to emerging usability of online information, the attention of effective result provisioning according to the user queries is an important process in Information Retrieval (IR) system. As exponential growth of a number of textual documents on the web, discovering the hidden information is often an arduous task in real-world applications. With the aim of addressing this constraint, Multiple Document Summarization (MDS) model creates a greater impact on the web searching field that facilitates the users to obtain the core information within a short time [1,2]. MDS focuses on providing the succinct and informative summaries from the larger document collections. This task is achieved by reducing the irrelevant sentences from the collection of documents to generate essential text summary in IR system. It is the extraction based MDS method that generates the summary by only omitting the sentences from the original text sentences not constructing the novel sentences. MDS is a non-trivial process in real time applications instead of summarizing a single document. For instance, News aggregation application in IR process: when multiple news feeds are submitted about a particular news topic, an IR system is forced to provide a short and comprehensive summary to users for improving the user convenience. It has the responsibility to create an understandable summary. Hence, an automatic multi-document summarization is necessary for IR system to summarize the multiple documents into an extractive summary.

To provide the accurate summarization, Ontology is the most valuable source that captures the hidden semantic information and comprises the abundant concepts and domain-related information [3]. It deals with either questions or input texts to identify the entities and its similarity based hierarchical structure. Most of the conventional researchers exploit the ontology to measure the semantic similarity [4] and to improve the document clustering [5] in text mining. Comparatively, some of the researchers are focused on the ontology-assisted document summarization. However, the extraction based MDS approaches to meet the redundancy problem since the top-ranked sentences of the multiple documents convey the similar information. Some of the existing methods resolve the redundancy issue while summarizing the multiple documents.

The main contribution of multi-document SummarY generatioN with tOPic ontology asSIStance (SYNOPSIS) approach includes two phases such as identification of sentence importance in the document set and optimal summarization of the document set.

- The SYNOPSIS approach summarizes the multiple documents under the same topic using Yago ontology and word position hypothesis, beneficial in effective IR process.
- The SYNOPSIS approach recognizes the entity of each keyword in the document sentences using Yago ontology after performing the preprocessing steps. Applying entity score, and word position, frequency and distance factors on the sentences enables the SYNOPSIS to assign the score to each sentence semantically.
- The SYNOPSIS approach reduces the sentence length by applying the structure analysis of the original sentences. Afterward, it creates the new scored sentence list by replacing the reduced sentences. It ensures that the context of the reduced sentence is as similar to the context of the original sentence. It

optimally selects the sentences to construct the final extractive summary, while satisfying the objective function.

- The experimental results show that the SYNOPSIS outperforms the conventional summarization method, since it balances the objective function of coherence, salience, and redundancy in the final summary.

1.1 Problem statement

As the documents in the MDS are relevant to the same topics, it encounters the similar text representations or contexts frequently in different documents. Thus, it results in the redundancy of the sentences which affects the quality of the extractive summary. MDS leads the several issues rather than using the single document summarization in which issues represent the compression, redundancy, passage selection, and speed. These are the crucial tasks while summarizing the collection of documents. Hence, the consideration of the redundant sentence reduction is often a challenging task in MDS system. Most of the existing summarization researchers provide the rank to the sentences in the documents without the knowledge of topic level information. Moreover, the several existing MDS systems lack in identifying the hidden semantic information and computing the most probable permutation of the original sentences since it does not exploit the ontology for document summarization. The earlier summarization methods contemplate the sentence similarity with the centroid, sentence length, and sentence position during the sentence selection process. In general, a sentence position hypothesis prefers the earlier sentence as the most important sentence than the other sentences in the document. It is not applicable to the real systems since the authors write the documents in their styles. Hence, the first sentence of the document may not be the relevant sentence in all the documents in which authors may write some other background information in the first sentence. The MDS system still faces several constraints such as providing concise ease of clear summary in a short time and balancing the coherence, salience, and redundancy in the summary. Hence, the proposed approach balances these objective functions in the MDS system.

2. Related work

The existing research works devote their significant efforts to summarize the collection of documents by using diverse data mining techniques.

2.1 Extraction based summarization

Extraction-based summarization models exploit the combination of statistical and linguistic features such as term frequency, sentence position [6], and topic signature [7] to discover the importance of the sentences. Text summarization performs the sentence extraction and compression process which employs the sub-tree based extraction [8] and integer linear programming in the constituent parser tree [9] respectively. Scaling Up MDS (SUMMA) [10] is a hierarchical summarization that enables the users to navigate a hierarchy of relatively short summaries in large scale. Ontology-enriched MDS (OMS) exploits background knowledge to improve summarization and maps the sentences with the ontology [11]. Graph-based approach [12] summarizes the documents based on the sentences in which the nodes refer the document sentences, and the weighted edges represent the similarity measure between the node-pair. The earlier works analyze the terms correlation in the documents using frequent itemset techniques [13], probabilistic approach [14], and Singular Value Decomposition (SVD) [15].

2.1.1 Lexical features

To identify the importance of a word in a document, the frequency measurement of a word in a document is significant. The sentence-level semantic analysis helps to determine the score of the sentence while summarizing the multiple documents [16]. Topic-focused MDS [17] employs the Relational Learning-to-rank (R-LTR) approach that avoids the diversity problem during summary extraction. A context sensitive document indexing scheme based on the Bernoulli model of randomness uses lexical association between phrases to determine the context sensitive weight [18].

2.1.2 Syntactic features

An approach [19] transforms Rhetorical Structure Theory (RST) trees into dependency trees using heuristic rules in single document summarization, formulates the summarization problem as the knapsack problem. The document summarization method [20] simplifies the approach in [19] by using the dependency-based discourse tree parser for directly providing discourse based summarization without any transformation. A nested tree structure based single document summarization exploits the rhetorical structure based sentence dependency and dependency parser based word dependency [21].

2.1.3 Machine learning

The machine learning techniques solve the complexity in feature selection when processing the different textual domains since the weight of the features is varied according to the domain. Hyper-graph [22] comprises the sentences and sentence relevance in which the graph-based ranking algorithm computes the sentence relevance on sub-topics. A standard Graph-based semi-supervised learning approach based on topic modeling strategies using two layers, i.e. Sentence layer and the topic layer improve the quality of the summary [23].

Most of the conventional summarization methods ranks the sentences based on the sentence position without considering the entity of the terms in the documents. It may affect the accuracy of the most succinct and comprehensive summary since unique writing style of the documents and topic distraction.

3. An Overview of SYNOPSIS methodology

With the aim of simplifying the reading complexity in the result of IR application, the SYNOPSIS approach is a target to summarize the concise and comprehensive information from the multiple documents. The conventional summarization techniques exploit the sentence position hypothesis that may select the irrelevant sentences in the actual context of the document since it assumes the first sentence in a document as the most significant sentence than the others. To tackle this constraint, the SYNOPSIS employs the word position hypothesis while ranking the sentences in the documents. The word position hypothesis identifies the importance value of a word in different sentences based on the appearances of the word. The SYNOPSIS approach incorporates two major phases such as identification of sentence importance in the document set, and optimal summarization of the document set.

Identification of Sentence Importance in the Document Set: The SYNOPSIS approach generates the entities of each feature or keyword based on the context of the keyword in each sentence using Yago ontology. Then, it provides the score to each sentence based on the entity score, word position and frequency in the document set, and word distance in the ontology structure. The sentence score only depends on the keywords among other than these keywords to identify the factual importance of the sentence in the documents.

Optimal Summarization of the Document Set: The SYNOPSIS reduces the sentence length and creates the duplicate sentences while ensuring the context of the document summarization using structural analysis. It generates the score for each duplicate sentence, i.e., reduced sentence, and replaces the original sentence by this duplicate sentence if reduced sentence has a higher score than the original sentence in the document. It optimally selects the high score-sentences and constructs the extractive summary based on the objective function that balances the coherence, salience, and redundancy.

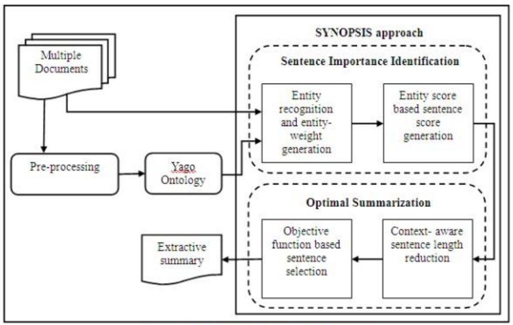

SYNOPSIS approach
Multiple Documents
Sentence Importance Identification
Entity recognition and entity-weight generation
Entity score based sentence score generation
Pre-processing
Yago Ontology
Optimal Summarization
Extractive summary
Objective function based sentence selection
Context- aware sentence length reduction

Figure 1: The SYNOPSIS methodology

3.1 Identification of Sentence Importance in the Document Set

The main goal of this phase is to discover the importance of the sentences in the whole document collections which facilitates the optimal summarization of the multiple documents. To achieve this goal, the SYNOPSIS employs the Yago Ontology to capture the hidden semantic knowledge of the sentences in the document in which ontology provides the entity of each feature or keyword in the sentences. The SYNOPSIS applies the preprocessing procedure to all the sentences in the document collections before capturing the entity based semantic knowledge from the ontology. Afterward, it generates the weight to the words based on the entity that is known as the weighted entity. Finally, it assigns the score to each sentence based on the entity score, word position and frequency in the document, and word distance in the Yago ontology structure.

3.1.1 Recognizing entity of the features

Initially, the SYNOPSIS performs the preprocessing steps on the sentences in the document set. The document set contains the related text information from the multiple documents, as the input, i.e. Sentences for the preprocessing procedure. The preprocessing steps include the sentence division, tokenization, stop words removal and stemming process. Preprocessing is the crucial venture process of evacuating the unnecessary words from the sentences that reduce the processing time of the system. The preprocessed sentence features are given as the input to the Yago ontology to explore the hidden information by recognizing the entity i.e. concept of each feature in the sentences. The ontology is the hierarchical structure of the source that contains the root node, a set of concepts, a set of is-a relations, a set of equivalent-class and a set of individuals. The utilization of ontology facilitates the MDS to understand the exact context of the documents.

Consider, a sentence (S_i^j) comprises the group of keywords (w_k) that are obtained from the preprocessing steps, where S_i^j represents the ith sentence in jth document ($d_j \in D_S$), and w_k represents a kth word. These feature or keyword set of each sentence is mapped to the Yago ontology hierarchy to identify the entities of the feature set. In ontology, a sentence is related to either one entity or two or more entities according to the nature of the features in the sentences. The SYNOPSIS recognizes the entities like a noun, date, time, and number from the Yago ontology. Also, the ontology provides the popularity score in terms of usage frequency, and related features for each entity. Each word in a sentence is mapped to the different entities since a word makes the different meaning in different contexts i.e. disambiguation. To identify the exact entity of each feature, according to the document context, the SYNOPSIS generates the weight to each entity based on the entity popularity and pertinent value to the document context. According to each word (w_k), the weighted entity (E_W) can be formulated as,

$$E_W(w_k) = \alpha * Pop(E_n) + \beta * Sim((con(E_n),con(d_j)) + (1-\alpha-\beta) * coh(E_n, D_S) \quad \text{.............}(1)$$

Where, α and β are the random variables that are user-specified parameters ϵ [0,1]. $Pop(E_n)$ is the popularity score of the entity of w_k, $Sim((con(E_n),con(d_j))$ is the similarity between the context of entity and context of the corresponding document, and $coh(E_n,D_S)$ is the coherence of the entity related to the document set.

3.1.2 Generating the score to each sentence

To discover the sentence importance, SYNOPSIS generates the score for each sentence in the documents based on the weighted entity and some essential factors.

The essential factors are a measurement of word position and word frequency in the document, and word distance in the ontology hierarchical structure. It is used to compute the sentence score (S_S) in the document that illustrates the significance of the document sentence. To calculate the sentence score, the SYNOPSIS computes the entity score based on the weighted entity in which entity may be either named entity or date, time and number entity. If an entity is a named entity ($N(E_n)$), the entity score is the addition of weighted entity and user-specified parameter (γ) which privileges the sentences. Otherwise, the entity score is only equal to the user-specified parameter. The sentence score is given as,

$$S_S(S_i^j) = \frac{\sum_{E_n \in E(S_i^j)} (\log F(w_k).P(w_k)) * (E_S(E_n)) * (1/D(w_k))}{|E(S_i^j)|} \quad \ldots\ldots\ldots\ldots(2)$$

Where, $|E(S_i^j)|$ is the set of entities of the ith sentence of the jth document. $F(w_k)$ and $P(w_k)$ represents the word frequency and word position in the document set respectively. ($E_S(E_n)$) and $D(w_k)$ refers the entity score and word distance in the ontology. The ontology based summarizer uses the recognized entity that provides the semantic knowledge to the sentences in the document collection. Equation (1) has been exploited to measure the entity score value that can be substituted in equation (2). The distance of the word depends on the feature level in the ontology structure in which the distance represents the number of levels between the entity and the corresponding word. The word position based sentence scoring depends on the earlier appearance of the particular word of a sentence as the most informative sentence than the other appearance sentences of that same word. Thus, this phase finally contains the scored sentences in the document that enables the optimal summarization phase by exploiting the sentences that have high sentence score.

3.2 Optimal Summarization of the Document Set

At the beginning of this phase, the SYNOPSIS aims at reducing the length of the sentences while ensuring the context relevant summary generation to make the succinct summary. To generate the concise and comprehensive summary, it creates the new list of sentences instead of having the original scored input sentence list. The new sentence list generation comprises only the reduced sentences of the high-rank sentences in which new sentence list is not a novel sentence formation. The reduced sentence is based on the words and sentence structure from the original sentence without deviating from the original concept of the sentence. Finally, it achieves the objective function that balances the coherence, salience, and redundancy in the extractive summary.

3.2.1 Reducing the sentence length

The SYNOPSIS exploits the high-rank sentences to reduce the length of the sentences by exploiting the sentence reduction algorithm. The sentence reduction algorithm employs the structure analysis that uses the six types of structures such as Adjectives, Appositions, Parentheticals, Adverbs or adverb phrases, Prepositional phrases, and Relative clauses. It reduces the sentence length in two cases by applying the structures onto the original sentences. Algorithm.1 shows the proposed algorithm.

In the first case, the SYNOPSIS contemplates the main clause based sentence reduction, if a sentence in the SVO structure. The sentence reduction algorithm identifies the main clause of the sentence and removes other than the main clause because the main clause is sufficient to understand the concept of the sentence. Hence, the main clause identification is beneficial to identify the removable structures in a sentence. In this sentence reduction, among six types of structures, five of them are used, namely appositions, adjectives, adverbs, parenthetical phrase, and relative clauses. These structures are considered to remove the unnecessary phrases from the main clause without changing its conceptual structure. As a result, the original sentence is directly replaced by this reduced sentence structure. For instance, ‘ In our country, only John got the Bronze medal’. From this sentence, the main clause is ‘only John got the Bronze medal’. In this main clause, the word ‘only’ is the adverb phrase, hence the removable phrase is ‘only’ while applying the structures in the main clause of the sentence.

In the second case, the SYNOPSIS employs the four structures, namely appositions, parenthetical phrases, relative clauses, and prepositional phrases, if a sentence is not in the SVO structure. This structure reduces the unnecessary phrases in the original sentence that results in the duplicate sentences. With the aim of reducing the sentence length, the SYNOPSIS generates the duplicate sentences using the words in the original sentence and removed phrases from a sentence. After that, it measures the sentence score to the duplicate sentences using equation (2) to compare the sentence score of the original sentence with the duplicate sentences. The sentence score shows the sentence relevance to the document context. Hence, the sentence score is calculated for the duplicate sentences, i.e., reduced sentences. If a duplicate sentence score ($S_S^d(S_i^j)$) is higher than the original sentence score ($S_S(S_i^j)$), the SYNOPSIS replaces the former sentence by the duplicate sentence. Since, it illustrates the highly informative sentence with a minimum number of words in a sentence. For instance, consider the original sentence has two removable structures, the SYNOPSIS creates the first duplicate sentence by removing those two removable structures. Then, it forms the second duplicate sentence by removing one of the removable structures and then, it constructs the third duplicate sentence by removing the only another one of the removable structure. Finally, it calculates the sentence score for these three duplicate sentences. These three sentences scores are evaluated with the original sentence score, which decides either replace it or not. If the original sentence is replaced by the duplicate sentence, the length of the summary can be reduced. Thus, this sentence reduction is beneficial to improve the effectiveness of the summary while ensuring the maximum informativeness of the final summary.

3.2.2 Generating the extractive summary

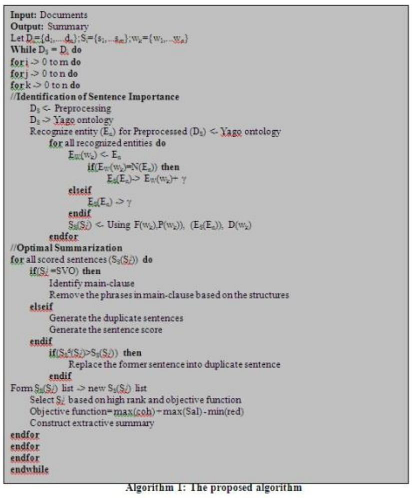
After the replacement of the reduced sentences, the SYNOPSIS retains the new sentence list that are nearly relevant to the document context. The SYNOPSIS selects the optimal sentences from all the scored sentences based on the objective function. If the sentences satisfy the objective function, the SYNOPSIS builds the extractive summary.

The objective function incorporates the coherence, salience, and redundancy. The coherence measurement depends on the relation between the sentences in the summary and salience measurement depends on the relation between the sentence and document context. The redundancy measurement depends on the relation between the sentences in the summary in which relation refers the similar information. An effective summary must have a high coherence and salience value and less redundant information regarding sentences. Hence, the proposed sentence selection depends on the coherence, salience, and redundancy. The MDS system contains the number of redundant sentences at the final stage since it incorporates the high-rank sentences of the multiple documents on the same topic or context. As a result, the redundant sentence removal is the crucial process in MDS system.

Initially, the SYNOPSIS removes the redundant information of sentences from the new sentence list i.e.final sentence list. To determine the hidden redundant information, it not only contemplates the ngrams of the root word but also considers the semantic relation of the root word. The semantic relation consists the synonyms, hypernyms, and hyponyms of the words in a document sentence. The complete semantic relations of the root word may distract the context of the sentence. Hence, the SYNOPSIS focuses on the threshold distance based semantic relation in the ontology tree structure. Equation (3) discovers the redundant information sentences that comprise only at the particular distance level from the root word. It contains the synonyms, hypernyms, and hyponyms of all the words in each sentence (S_m) as the ngrams that is mapped to the ngrams of the other sentences (X) in the document.

$$Sim(S_m,X)= \frac{1}{n}\sum_{k=1}^{n}\frac{|ngrams(S_m) \cap ngrams(X)|}{ngrams(X)} \quad \ldots\ldots\ldots\ldots\ldots\ldots(3)$$

Where, ‘m’ and ‘n’ represents the number of sentences in the document and number of words present in a sentence respectively. According to equation (3), the SYNOPSIS computes the similarity between sentence by mapping each sentence with a sentence from the other sentences set. This semantic similarity facilitates the redundancy based sentence removal that is the high semantic similarity of the sentence pair have the redundant information. Thus, the SYNOPSIS removes the redundant sentences based on the threshold value. After removing the redundant sentences, the SYNOPSIS examines the retained sentences to form the extractive summary.



```
Input: Documents
Output: Summary
Let D_i={d_1,...d_n};S_i={s_1,...s_m};w_k={w_1,...w_n}
While D_S = D_i do
for i -> 0 to m do
for j -> 0 to n do
for k -> 0 to n do
//Identification of Sentence Importance
    D_S <- Preprocessing
    D_S -> Yago ontology
    Recognize entity (E_a) for Preprocessed (D_S) <- Yago ontology
        for all recognized entities do
            E_W(w_k) <- E_a
                if(E_W(w_k)=N(E_a)) then
                    E_S(E_a)-> E_W(w_k)+ γ
            elseif
                E_S(E_a) -> γ
            endif
            S_S(S_i') <- Using F(w_k),P(w_k)), (E_S(E_a)), D(w_k)
        endfor
//Optimal Summarization
for all scored sentences (S_S(S_i')) do
    if(S_i' =SVO) then
        Identify main-clause
        Remove the phrases in main-clause based on the structures
    elseif
        Generate the duplicate sentences
        Generate the sentence score
    endif
        if(S_S^d(S_i')>S_S(S_i')) then
            Replace the former sentence into duplicate sentence
        endif
Form S_S(S_i') list -> new S_S(S_i') list
    Select S_i' based on high rank and objective function
    Objective function=max(coh)+max(Sal)-min(red)
    Construct extractive summary
endfor
endfor
endfor
endwhile
```

Algorithm 1: The proposed algorithm

The second factor of the objective function is salience. The SYNOPSIS examines the salience of the set of sentences list of the extractive summary that exemplifies the amount of context-sensitive sentences in the final summary. Usually, the proposed sentence score based sentence selection offers the optimal sentence, i.e., high salience sentence to the final summary. Also, it focuses on the salience measurement of each sentence. It is derived from the equation (3) by modifying the factors (S_m,X) into (D_S,S_m) in which ‘D_S’ refers the document set. It examines the presence of the words in a sentence with the presence of the words in the document set to identify the inherent value of each sentence. The high salience sentences get the first preference in the sentence ordering while ensuring the coherence.

Finally, the SYNOPSIS focuses on both coherent sentence selection from the retained sentence list and coherent sentence ordering. The goal of this process is to discover the pairwise ordering sentences among the overall sentences in the document based on the relation link between the sentences. It employs the appearance of the words in both the sentences to determine the edge weight. If edge weight between the sentence is high, these sentences have consistently related information in which the edge weight is measured by using the discourse link of ontology structure. The SYNOPSIS builds the extractive summary based on the high coherence, salience, and less redundancy of the sentences. Thus, the SYNOPSIS balances the objective function while ensuring concise and comprehensive MDS.

4. Experimental Evaluation

This paper evaluates the SYNOPSIS approach with the R-LTR approach [17] to exemplify better performance of the SYNOPSIS approach than the conventional summarization method.

4.1 Experimental setup

The SYNOPSIS approach is implemented by exploiting the Java platform and an expert system of Java Expert System Shell (JESS) rule engine. In this experiment, it removes the stop words and applies stemming from the remaining words of the document sentences using Porter stemmer. This approach exploits Yago ontology to provide entity, the semantic relation of each keyword and uses the WEKA machine learning toolkit to obtain the best results. The proposed implementation employs Recall-Oriented Understudy for Gisting Evaluation (ROGUE) which is the automatic summarization evaluation package to examine the SYNOPSIS. The ROUGE shows the quality of the final summary based on the n-gram word measurement. It evaluates the author made catchphrases i.e., reference summary with the system generated catchphrases i.e.final summary.

4.1.1 Dataset

The SYNOPSIS approach uses the documents from the UCI machine learning repository. The text corpus includes 4000 legal cases of Australian legal cases from the Federal Court of Australia (FCA). The data is gathered from the year of 2006, 2007, 2008, and 2009. All the documents are segmented into sentences that contain both stop words and keywords. The Australasian Legal Information Institute provides the author made catchphrases evaluate the performance of the proposed summary.

4.1.2 Evaluation metrics

Precision: Precision is the ratio between the number of sentences occurring in both final summary and reference summary and the number of sentences in the final summary.

Recall: Recall is the ratio between the number of sentences occurring in both final summary and reference summary and the number of sentences in the reference summary.

F-measure: F-measure is the composite measure of precision and recall combination.

$$\text{F-measure} = 2 * \left(\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \right)$$

.2 Experimental results

4.2.1 Precision

Fig.2 shows the precision of both SYNOPSIS and R-LTR approach while varying the length of the summary as well as the Average Number of Sentences (ANS). The ANS is referred as the average number of sentences in the input document set. The SYNOPSIS approach constructs the summary with the maximum length of 25 sentences. The precision value increases when increasing the summary length that is when reaching the maximum number of sentences in the summary. The precision value of SYNOPSIS approach is suddenly increased by 9.63% while varying the summary length from 5 sentences to 25 sentences at the point of ANS=50. In the same scenario, the R-LTR approach has marginally increased by only 7.99%. This development is achieved by exploiting the word position hypothesis, word distance, word frequency, and entity score based sentence ranking in MDS system. When the summary reaches the maximum length, i.e., 25 sentences and ANS=150, the R-LTR approach has lagged 3.12% while comparing with the SYNOPSIS approach.

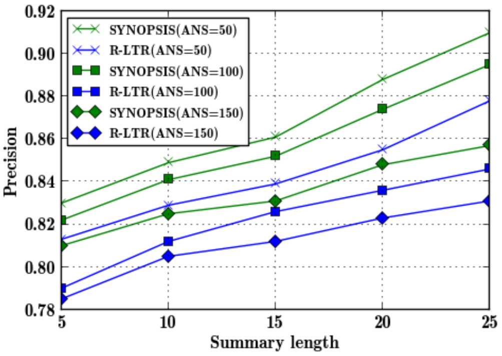


Figure 2: Precision of SYNOPSIS

4.2.2 Recall

The recall of both the SYNOPSIS and R-LTR approach is illustrated in Fig.3. It shows the proposed system accuracy when changing the number of input sentences and final summary length. The recall of the SYNOPSIS approach is increased by 2% while comparing the R-LTR approach if ANS=50 and summary length reaches the maximum level since the SYNOPSIS approach reduces sentence length by generating the context sensitive duplicate sentences. It also contemplates the objective function of coherence, salience, and redundancy based sentence selection and ordering in the final summary. The performance in terms of accuracy of the SYNOPSIS approach at ANS=150 is nearly obtained in the performance of existing R-LTR approach at ANS=50. The recall of the R-LTR approach has marginally increased by 6.96% when summary length increased from 5 sentences to 25 sentences in the case of ANS=150.

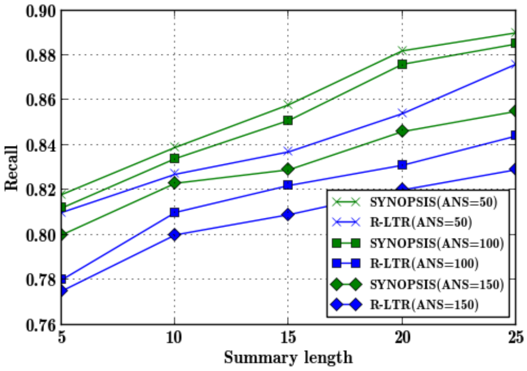


Figure 2: Recall of SYNOPSIS

4.2.3 F-measure

Fig.4 indicates the F-measure of both the SYNOPSIS and R-LTR approach while varying the redundancy ratio and Salience. Redundancy ratio is the ratio between the number of redundant sentences and the total number of sentences shows the amount of redundant information in the input document. F-measure of the SYNOPSIS approach is balanced by the particular amount of redundant information in the document set, after that it marginally decreases the performance. However, the R-LTR approach cannot maintain the balanced F-measure value, it suddenly decreases the performance when escalating the redundancy ratio. If the salience of the specific sentence is 20% and the redundancy value of that same sentence is 0.5, the SYNOPSIS approach has increased by 9.33% comparatively. Since the SYNOPSIS approach semantically removes the redundant information using Yago ontology. It reduces the redundant sentences before validating the salience of the sentences. Hence, it improves the F-measure when high redundant information of that sentence has high salience.

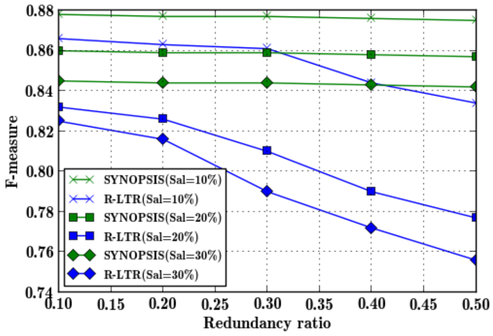


Figure 4: F-measure of SYNOPSIS

4.2.4 Recall vs. Compression factor

Fig.5 illustrates the recall of both the SYNOPSIS and R-LTR approach when escalating the compression factor and an average number of words (ANW). The compression factor is depending on the number of words in the final summary and an average number of words in the document set. The R-LTR approach initially increases and then decreases when the compression factor is increased but, the SYNOPSIS approach marginally increases the recall value due to the consideration of structure analysis based context-aware summary generation. The R-LTR approach has degraded by 10.59% while comparing the SYNOPSIS approach when the compression factor is 0.8 and ANW=500, since, the SYNOPSIS coherently selects and orders the sentences while ensuring the non-redundancy in the final summary.

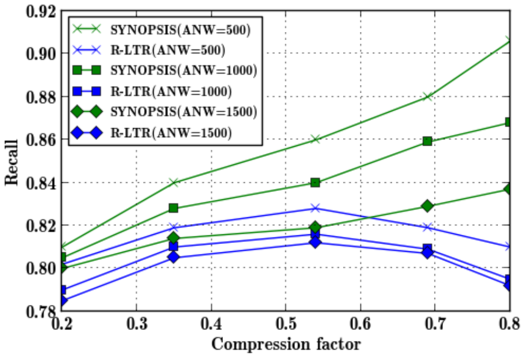


Figure 5: Recall vs. Compression factor

5. Conclusion

This paper investigates the SYNOPSIS approach to coherently generate the concise and comprehensive summary by focusing on optimal sentence ranking and sentence selection. The main goal of this SYNOPSIS approach is achieved by integrating an ontology and a word position hypothesis based sentence evaluation and selection process. The SYNOPSIS employs the Yago ontology to score the sentences based on the entity of the keywords, and word position, frequency, and distance semantically. To provide the useful summary, it further reduces the length of the sentences without modifying the concept of the sentence by applying structure analysis. It generates the score to reduced sentences and creates the new sentence list for the extractive summary. Finally, it selects the high-rank sentences while satisfying the objective function. The experimental results show the high accuracy of IR process in MDS system.

References

[1] Das, Dipanjan, and André FT Martins, "A survey on automatic text summarization", Literature Survey for the Language and Statistics, Vol.4, pp.192-195, 2007

[2] Ding Yuan, "A Survey on Multi-Document Summarization", Department of Computer and Information Science University of Pennsylvania, 2004

[3] Wu, Keshou, Lei Li, Jingxuan Li, and Tao Li, "Ontology-enriched multi-document summarization in disaster management using submodular function", Information Sciences, Vol.224, pp.118-129, 2013

[4] D. Sachez, M. Batet, A. Valls, and K. Gibert, "Ontology-driven web-based semantic similarity", Springer transaction on Intelligent Information Systems, 2009

[5] Hotho, Andreas, Steffen Staab, and Gerd Stumme, "Ontologies improve text document clustering", Third IEEE International Conference on Data Mining, pp. 541-544, 2003

[6] R. Katragadda, P. Pingali, and V. Varma, "Sentence Position Revisited: A Robust Light-Weight Update Summarization 'Baseline' Algorithm", ACM Proceedings of the Third International Workshop on

Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies, pp.46-52, 2009

[7] Sizov Gleb, "Extraction-Based Automatic Summarization: Theoretical and Empirical Investigation of Summarization Techniques", 2010

[8] Hajime Morita, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura, "Subtree Extractive Summarization via Submodular Maximization", Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp.1023–1032, 2013

[9] Chen Li, Yang Liu, Fei Liu, Lin Zhao, and FuliangWeng, "Improving Multi-documents Summarization by Sentence Compression based on Expanded Constituent Parse Trees", Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.691-701, 2014

[10] Janara Christensen, Stephen Soderland, Gagan Bansal, and Mausam, "Hierarchical Summarization: Scaling Up Multi-Document Summarization", Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp.902-912, 2014

[11] Lei Li, Dingding Wang, Chao Shen, and Tao Li, "Ontology-enriched Multi-Document Summarization in Disaster Management", Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp.819-820, 2010

[12] Thakkar, Khushboo S., Rajiv V. Dharaskar, and M. B. Chandak, "Graph-based algorithms for text summarization", IEEE 3rd International Conference on Emerging Trends in Engineering and Technology (ICETET), pp.516-519, 2010

[13] Baralis, E., Cagliero, L., Fiori, A., and Jabeen, S, "Multi-document summarization exploiting frequent itemsets", ACM Proceedings of the Symposium on Applied Computing, pp.782-786, 2012

[14] Conroy, J., Schlesinger, J., Kubina, J., Rankel, P., and OLeary, D, "Guided and multi-lingual summaries and evaluation metrics", In TAC: Proceedings of the Text Analysis Conference, 2011

[15] Steinberger, J., Kabadjov, M., Steinberger, R., Tanev, H., Turchi, M., and Zavarella, V, "Guided and multilingual summarization tasks", In TAC'11: Proceedings of the Text Analysis Conference, 2011

[16] DingdingWang, Tao Li, Shenghuo Zhu, and Chris Ding, "Multidocument summarization via sentence-level semantic analysis and symmetric matrix factorization", ACM Proceedings of the 31st annual international SIGIR conference on Research and development in information retrieval, pp.307–314, 2008

[17] Zhu, Yadong, Yanyan Lan, Jiafeng Guo, Pan Du, and Xueqi Cheng, "A novel relational learning-to-rank approach for topic-focused multi-document summarization", IEEE 13th International Conference on Data Mining (ICDM), pp.927-936, 2013

[18] Pawan Goyal, Laxmidhar Behera, and Thomas Martin McGinnity, "A Context-Based Word Indexing Model for Document Summarization", IEEE transactions on Knowledge and Data Engineering, Vol.25, No.8, pp.1693-1705, 2013

[19] Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata, "Single-Document Summarization as a Tree Knapsack Problem", Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (ACL), pp.1515-1520, 2013

[20] Yasuhisa Yoshida, Jun Suzuki, Tsutomu Hirao, and Masaaki Nagata, "Dependency-based Discourse Parser for Single-Document Summarization", Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics (ACL), pp.1834-1839, 2014

[21] Yuta Kikuchi, Tsutomu Hirao, Hiroya Takamura, Manabu Okumura Masaaki Nagata, "Single Document Summarization based on Nested Tree Structure", Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL) (Short Papers), pp.315- 320, 2014

[22] Wei Wang, Furu Wei, Wenjie Li, and Sujian Li, "Hypersum: hypergraph based semi-supervised sentence ranking for query oriented summarization", ACM Proceedings of the 18th conference on Information and knowledge management, pp.1855-1858, 2009

[23] Yanran Li, and Sujian Li, "Query-focused Multi-Document Summarization: Combining a Topic Model with Graph-based Semi-supervised Learning", Proceedings of COLING, 25th International Conference on Computational Linguistics: Technical Papers, pp.1197-1207, 2014