American Journal of Applied Sciences 9 (7): 1063-1070, 2012
ISSN 1546-9239

Sentence Annotation based Enhanced Semantic Summary Generation from Multiple Documents

Kogilavani, A. and P. Balasubramanie
Department of Computer Science and Engineering,
Kongu Engineering College, Erode, India

Abstract: Problem statement: The goal of document summarization is to provide a summary or outline of manifold documents with reduction in time. Sentence extraction could be a technique that is employed to pick out relevant and vital sentences from documents and presented as a summary. So there is a need to develop more meaningful sentence selection strategy so as to extract most significant sentences. **Approach:** This study proposes an approach of generating initial and update summary by performing sentence level semantic analysis. In order to select the necessary information from documents all the sentences are annotated with aspects, prepositions and named entities. To detect most dominant concepts within a document, Wikipedia is used as a resource and the weight of each word is calculated using Term Synonym Concept Frequency-Inverse Sentence Frequency (TSCF-ISF) measure. Sentences are ranked based on the scores they have been assigned and the summary is formed from the highest ranking sentences. **Results:** To evaluate the quality of a summary based on coverage between machine summary and human summary intrinsic measures called Precision and Recall are used. Precision is used to determine exactness whereas Recall is used to measure the completeness of the summary. Then our results are compared with LexRank Update summarization task and with the Semantic Summary Generation method. The ROUGE-1 measure is used to identify how well machine generated summary correlates with human summary. **Conclusion:** The performance of update summarization relies highly on measurement of sentence similarity based on TSCF-ISF. The experiment result shows that low overlap between initial summary and its update summary.

Key words: Term Synonym Concept Frequency-Inverse Sentence Frequency (TSCF-ISF), sentence annotation, semantic element extraction, sentence scoring, initial summary, update summary

INTRODUCTION

Recently, online web content data are raised in an increasing speed, people should develop a crisp overview from a large number of articles in a tiny point in time. So document summarization, aim at generating concise, comprehensible and semantically meaningful summaries. Multiple document summarization aims at extract most vital information from several documents. Producing updated information could be a valuable technique for people to urge latest information by eliminating surplus data. The aim of multi-document update summary generation is to construct a summary unfolding the mainstream of data from a collection of documents with the hypothesis that the user has already read a set of previous documents. This sort of summarization has been proved significantly helpful in tracing news stories, solely new data got to be summarized if we had previously known a little about the story. In order to provide a lot of semantic information, guided summarization task is introduced by the Text Analysis Conference (TAC). It aims to produce semantic summary by using a list of important aspects. The list of aspects defines what counts as important information but the summary also includes other facts which are considered as especially important. Furthermore, an update summary is additionally created from a collection of later Newswire articles for the topic under the hypothesis that the user has already read the previous articles. The summary generated is guided by pre-defined aspects that is employed to enhance the quality and readability of the resulting summary.

Using term frequency to determine important concepts in a text has proven to be successful because of its simplicity and universal applicability, but statistical methods can only provide the most basic level of performance. To address this issue the

Corresponding Author: Kogilavani, A., Department of Computer Science and Engineering, Kongu Engineering College, Erode, India

proposed system employs term synonym concept frequency-inverse sentence frequency measure. In order to produce a responsive summary meaning oriented structural analysis (Jin *et al*., 2011) is needed. To address this issue the proposed system presents a document summarization approach based on sentence annotation with aspects, prepositions, named entities. Semantic element extraction strategy is used to select important concepts from documents which is used to generate an enhanced semantic summary. Extensive experiments on the TAC 2008 datasets illustrate that the proposed method outperforms the state-of-the-art system.

Background: Developed Wikipedia-based summarization system WikiSummarizer which discusses about sentence wikification, i.e., Enriching sentence representation with concepts from Wikipedia. Also, semantic relatedness of Wikipedia concepts are considered to produce a summary. But other forms of information in Wikipedia are needs to be examined for creating a more comprehensive representation of sentences. Kogilavani and Balasubramanie (2011a) developed a semantic summary by constructing semantic vector space model with dependency parse relations which utilizes action words. Relevant sentences are selected by applying different combinations of features. The main drawback of this approach is that there is no precise information structure. Barrera and Verma (2010) developed a ranking-based approach which introduces a prioritization hierarchy consisting of four levels that are used to determine the most important sentences for extraction. Level 1 considers a sentence's distinct types of entities count. Level 2 utilizes an article level rank based on article date. Level 3 is based on the normalized score based on sentence's total entity count. Level 4 is based on syntactic, semantic and statistical methodologies. Sentences with more types of names entities and total entities give the summary a better linguistic quality. In this approach further investigation is needed to eliminate Level 3 tiebreaking method or reversal of Levels 3 and 4. Varma *et al*. (2010) developed a summarization system with knowledge based measures and utilized domain and sentence tag models to score sentences. Since the focus is on guided summarization, this method resulted in poor performance. Long *et al*. (2010) developed a new method for update summary generation which utilizes morphological features of a sentence. According to this approach sentences with diverse essential elements are selected. But to create a good summary a heuristic method will be required. The PSO was employed in Binwahlan *et al*. (2009; 2010) to calculate the weight of the text features. This is done to get the best text features. In order to calculate the score for each sentence the fuzzy inference system was used.

Kumar and Salim (2011) various surveys on multiple document summarization approaches has been offered. This study discusses about feature, cluster, graph and knowledge based methods for summary generation.

MATERIALS AND METHODS

The proposed approach to generate semantically enhanced initial and update summary from multiple documents is shown in Fig. 1.

A collection of topic related two sets of documents are fed as input. The output is a concise set of two summaries that contains reduced information. The main aim is to simulate a user who is interested in learning about the latest developments on a specific topic and who wishes to read a brief summary of the latest news. The proposed method can be split into the following modules: (1) summary generation algorithm (2) sentence annotation (3) Wikipedia based semantic element extraction (4) initial summary generation (5) update summary generation.

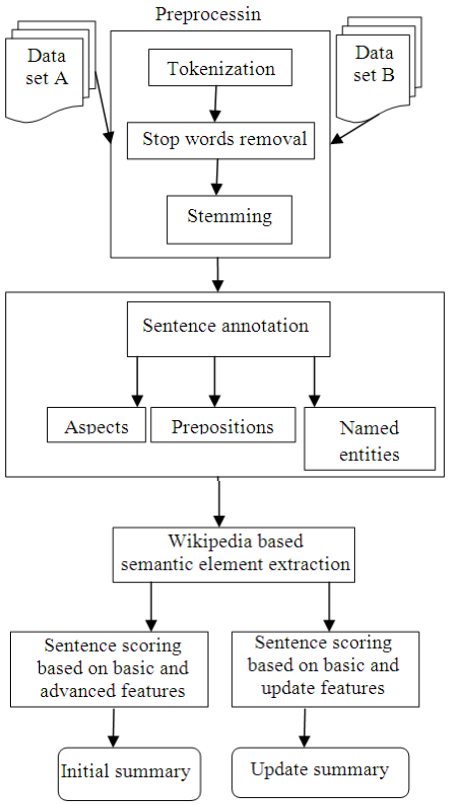


Fig. 1: Proposed system model

The next explosion occurred at 8:56 a.m. near King's Cross Station, where the death toll was 21, the police said. Twenty-one minutes later, at 9:17 a.m., a third blast ripped through a train coming in Edgware Road underground station, killing seven. Annotation of the sentence with above mentioned aspects are given below.

(a)

The next explosion occurred at <when>8:56 a.m.</when> near <where>King's Cross Station</where>, where the <who affected>death toll was 21<who affected>, the police said. Twenty-one minutes later, at <when>9:17 a.m. </when>, a third blast ripped through a train coming in <where>Edgware Road</where> underground station, <who affected>killing seven<who affected>.

(b)

The <preposition>next</preposition> explosion occurred <preposition>at</preposition> 8:56 a.m. <preposition>near</preposition> King's Cross Station, where the death toll was 21, the police said. Twenty-one minutes later, at 9:17 a.m., a third blast ripped <preposition>through</preposition> a train coming <preposition>in</preposition> Edgware Road underground station, killing seven.

(c)

The next explosion occurred at <time>8:56 a.m.</time> near <location>King's Cross Station</location>, where the death toll was 21, the <person name>police</person name> said. Twenty-one minutes later, at <time>9:17 a.m.</time>, a third blast ripped through a train coming in <location>Edware Road</location> underground station, killing seven.

(d)

Fig. 2: (a) Sample sentence (b) Sentence annotated with aspects (c) Sentence annotated with prepositions (d) Sentence annotated with named entities

Summary generation algorithm:

Step 1: Initially the articles in the dataset are split into sentences and those sentences are annotated with predefined aspects, prepositions and Named entities.

Step 2: Sentence representation is enhanced by extracting concepts from Wikipedia, which is referred to as a sentence unification process.

Step 3: Individual sentences are mapped into concepts and individual word score is calculated based on novel TSCF-ISF measure.

Step 4: Then for each sentence, score is calculated based on Basic and Advanced features for dataset A articles and based on Basic as well as Update features for dataset B articles.

Step 5:Highest ranking sentences are selected and ordered in a way in which the sentences are included in the original documents and final initial summary is generated.

Step 6: Update summary is generated after removing redundancy.

Sentence annotation with aspects: The articles from datasets are split into sentences and annotated with appropriate template tags. These annotations include both objective (when, where, who) and subjective (how, why, countermeasures) tags (Owczarzak and Dang, 2011). As any standard Named Entity Recognition can only tag objective tags, we chose to manually annotate all the articles with all possible tags. A sentence is tagged with multiple tags it has more than one answer to the template. For example consider the following sentence taken from the document D08021D:NYT_ENG_20050707 related to Attacks category. Figure 2a denotes sample sentence and Fig. 2b denotes sentence with aspects.

Sentence annotation with prepositions: In English grammar, a preposition is a part of speech that links nouns, pronouns to other phrases in a sentence. A preposition generally represents the temporal, spatial or logical relationship of its object to the rest of the sentence. It is very interesting to observe how prepositions are implicitly capturing the key elements in a sentence. The list of prepositions used for calculating sentence importance are limited to simple single word prepositions like in, on, of, at, for, from, to, by, with. Annotation of the above sentence with prepositions are given in Fig. 2c.

Sentence annotation with named entities: Prior observations in the given data led to believe that more the types of names entities a sentence contains, the stronger the likelihood the sentence's capabilities are in answering a set of questions like what happened? Who was involved? And where did this happen? Named entities refer to the objects for which proper nouns are used in a sentence. Seven basic named entities are identified: person, location, date, time, organization, money and percentage. Stanford Named Entity Recognition (NER) is employed to identify person, location, organization entities. Others are extracted by applying patterns. Annotations of the above sentence with named entities are given in Fig. 2d.

Wikipedia based semantic element extraction: Words are conventionally considered to be the units of text to calculate importance. Simple word counts and frequencies and synonym based word frequencies in the document collection have proved to work well in the context of summarization. The proposed system uses semantic concepts in computing sentence importance. Wikipedia is a vast, interlinked articles providing a multilingual database of concepts, web-based, free-content encyclopedia, comprehensive and well-organized knowledge repository. The links are there in Wikipedia articles which is used to direct the user to recognize related pages. Wikipedia Miner is a freely available toolkit for navigating and making use of content of Wikipedia. The proposed system creates concept database from Wikipedia concepts by selecting the concepts that appear explicitly in a sentence and each word in each sentence is compared with concept database.

Let D = {d_1, d_2, d_3…… dk) be the set of documents where k is the number of documents in D. Let N = {s_1, s_2, s_3…… Sn} be the number of sentences in D which can be calculated during preprocessing. Let M = {w_1, w_2, w_3…… Wm} be the number of words in each sentence after removing stop words. Let C = {c1, c2,… can} be the set of concepts in the concept database. Let d_i be the i^{th} document in D, $S_{i,k}$ be the i^{th} sentence in any document $d_{k,}$ w_m be a word in a sentence $S_{i,k}$. To improve accuracy and to calculate the weight of each word, the proposed system adopts Term Synonym Concept Frequency (TSCF). Every word's TSCF is calculated by performing synset extraction, Concept Database construction and term frequency calculation. The Term Synonym Concept Frequency (TSCF) of every word is obtained by Eq. 1:

$$TSCF(w_i) = \sum_{w_i \in \{\{w\} \cup synonym(w)\}} \alpha.TF(w_i) + \beta \quad (1)$$

In TSCF calculation to include word synonym into account the Tern Frequency (TF) of each word and its synonym is multiplied by α where α = 1 for the word and α = 0.5 for synonym of the word and β = 1 if the word itself is a concept in the concept database. Synonym is retrieved from WordNet, a lexical database for the English language. The Term Frequency (TF) of each word is calculated according to Eq. 2 (Kogilavani and Balasubramanie, 2011a):

$$TF(w_m) = \frac{n_m}{\sum_k n_k} \quad (2)$$

where, n_m is the count of the m^{th} word appears in D. For example if word 'cargo' occurs 10 times in document collection D, then n_m value is 10. This value is divided by the number of occurrences of all words in all sentences of D. Inverse sentence frequency is calculated as Eq. 3:

$$ISF(w_m) = \log \frac{N}{S} \quad (3)$$

where, S is the count of sentences that contain m^{th} word. Then for each sentence the importance of words in that sentence will be calculated by TSCF*ISF value.

Initial summary generation: To generate initial summary or general summary, there is a need to capture the relevant sentences from multiple documents. Relevant sentences are selected based on different features. The proposed work combines six features from (Kogilavani and Balasubramanie, 2011b) which is referred to as basic features with new additional features referred to as advanced features like sentence annotation with aspects, prepositions, named entities and sentences with semantic concepts feature. During initial summary generation, a subset of rank sentences is selected to generate a summary. A redundancy check is done between a sentence and summary generated so far, before selecting it in the summary. Sentences are adjusted on their order of occurrence in the original documents to improve readability.

Basic Feature 1 word-feature: The significance of each word is calculated by using a novel measure Term Synonym Concept Frequency-Inverse Sentence Frequency (TSCF-ISF) Eq. 4:

$$W\_F(s_{i,k}) = \sum Word\_Score(s_{i,k}).f(w_m, s_{i,k}) \quad (4)$$

where, $f(w_m, s_{i,k})$ is the frequency of each word w in sentence $s_{i,k}$ Eq. 5:

$$Word\_Score(s_{i,k}) = \sum_{i=1}^{m} TSCF(w_i).ISF(w_i) \quad (5)$$

Remaining Basic Features 2-6 are selected from (Kogilavani and Balasubramanie, 2011b).

Advanced Feature 1 sentence annotation with aspects: Any sentence that contains important aspects are considered as an important one. This feature is calculated as Eq. 6:

$$A\text{-}F(S_{i,k}) = \frac{A\_Count(S_{i,k})}{Length(S_{i,k})} \quad (6)$$

where, A-Count ($S_{i,k}$) is a count of annotations in $S_{i,k}$.

Advanced Feature 2. SentenceAnnotation with a preposition: A sentence is considered as important one if it consists of more number of prepositions. Hence this feature is calculated as Eq. 7:

$$\text{Pre-F}(S_{i,k}) = \frac{\text{Pre\_Count}(S_{i,k})}{\text{Length}(S_{i,k})} \quad (7)$$

where, Pre_Count($S_{i,k}$) is a count of prepositions in $S_{i,k}$.

Advanced Feature 3 sentence annotation with named entities: A sentence with more Named Entities are important ones. Hence this feature is calculated as Eq. 8:

$$\text{NE\_F}(S_{i,k}) = \frac{\text{NE\_Count}(S_{i,k})}{\text{Length}(S_{i,k})} \quad (8)$$

where, NE_Count ($S_{i,k}$) is a count of Named Entities in $S_{i,k}$.

Advanced Feature 4 sentences with semantic concepts: If a sentence has more number of semantic concepts then it is considered as salient one. This feature is calculated as Eq. 9:

$$\text{SC\_F}(S_{i,k}) = \frac{\text{SC\_Count}(S_{i,k})}{\text{Length}(S_{i,k})} \quad (9)$$

where, SC_Count ($S_{i,k}$) is a count of semantic concepts in a sentence $S_{i,k}$.

The score of each sentence is calculated using Eq. 1-9 by considering only Basic Features and Basic Features with Advanced Feature1, Basic Features with Advaned Feature 2, Basic Features with Advaned Feature 3, Basic Features with Advaned Feature 4 and finally all Basic Features with All Advanced Features. Initial summary is generated by taking highest scoring sentences.

Update summary generation: To generate update summary six Basic Features and three Update specific features are used. Two Update features are defined in (Kogilavani and Balasubramanie, 2011a) and third feature is defined as follows.

Update Feature 3 Novel Sentence Similarity Measure (NSSM): This new feature selects novel sentences that have not been contained in the initial summary. All sentences in initial summaries are considered as candidate sentences. New sentences that have least similarity with these candidate sentences are chosen as sentences in update summary. The similarity between candidate sentences and sentences in dataset B is calculated as follows Eq. 10:

$$\text{Sim}(S1,S2) = \frac{\sum w_i}{\sum w_j} \quad (10)$$

where, $w_i \in S1 \cap S2$, $w_j \in S_{min}$. The numerator is the sum weight of the words that both occur in sentence s1 and s2. The denominator is the sum weight of the words that in the short sentence S_{min} in {s1, s2}.

The benefit is that if a sentence contains all the words of another sentence, i.e. If one sentence is totally a part of another, then their similarity is 1.

RESULTS AND DISCUSSION

The proposed summarization approach will be evaluated on the TAC 2008 dataset. Firstly the datasets and evaluation criteria are introduced as follows.

Dataset: The dataset from text analysis conference 2008 were used in our experiments. This dataset called as AQUAINT-2 corpus consists of news articles from October 2004 to March 2006. Dataset consists of 48 topics, 20 documents per topic in chronological order. The entire dataset is arranged into two clusters of articles, referred to as dataset A and B in which B articles were more recent than dataset A articles and the summary of the second cluster had to provide only an update about the topic, avoiding any repetition of information from the first cluster. The main task in the proposed system is to produce guided and semantically enhanced initial summary of a set of an article. Update task is to produce update summary from a collection of B articles by assuming that the information in the first set is already known to the reader.

Evaluation criteria: We evaluated our method by comparing the generated summaries to human summaries under three different measures like precision, recall and ROUGE-1 measure. To evaluate the quality of a summary based on coverage between machine summary and human summary an intrinsic measure called Precision and Recall measures are used. Then our results are compared with LexRank Update summarization task and with the semantic summary generation method. The ROUGE-1 measure is used to identify how well automated summary correlates with summary generated manually.

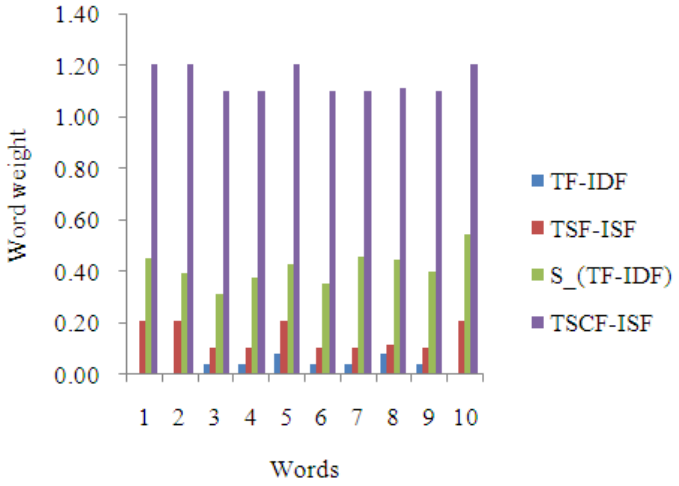
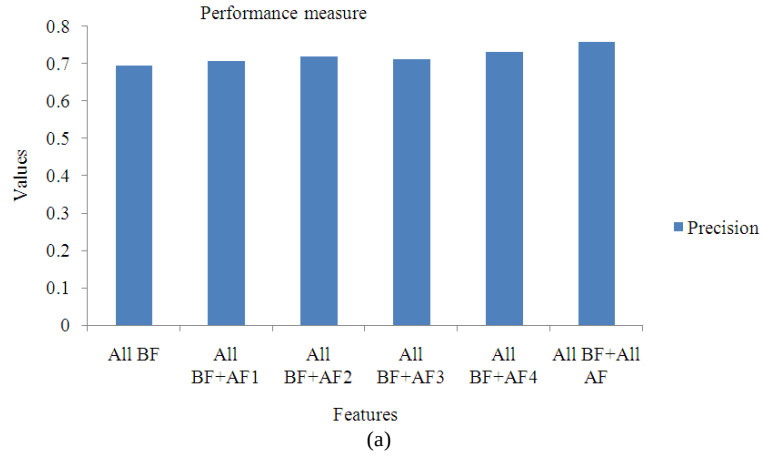
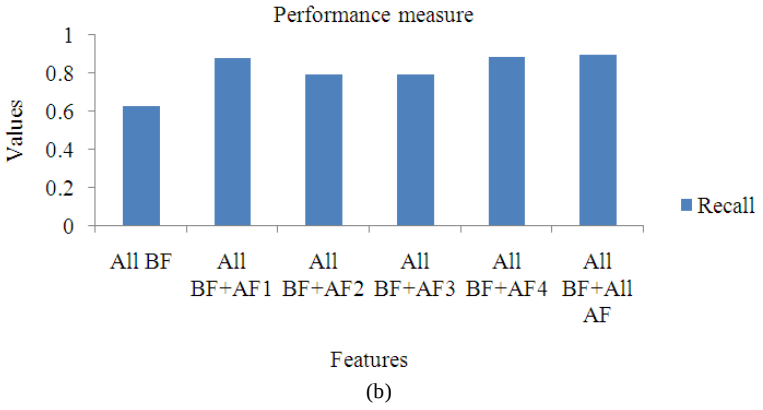


Fig. 3: Comparison between measures



(a)



(b)

Fig. 4: (a) Initial summary-precision (b) Initial summary-recall

Figure 3 shows word score calculated by TF-IDF, TSF-ISF, S_(TF-IDF), TSCF-ISF. The result indicates that improved accuracy is obtained by TSCF-ISF measure. Figure 4a and b represents the performance

measure based on precision and recall for all six Basic Features (BF), Six Basic Features combined with Advanced Feature1 (BF+AF1), Six Basic Features combined with Advanced Feature2 (BF+AF2), Six Basic Features combined with Advanced Feature3 (BF+AF3), Six Basic Features combined with Advanced Feature4 (BF+AF4), Six Basic Features combined with all advanced Features (BF + All AF).

The chart shows that when basic features are combined with all Advanced Features, the precision and recall is high compared to all other feature combinations. By incorporating sentence specific features along with TSCF-ISF, the precision is improved which implies that the coverage and completeness in machine summary is improved.

Figure 5a and b represents the performance measure based on precision and recall for all six Basic Features (BF) combined with Update Feature1 (BF+UF1), Six Basic Features combined with Update Feature2 (BF+UF2), Six Basic Features combined with Update Feature3 (BF+UF3), Six Basic Features combined with all three Update Features (BF+UF1+UF2+UF3). The chart shows that when considering all Update Features, the precision and recall is high compared to all other feature combinations.

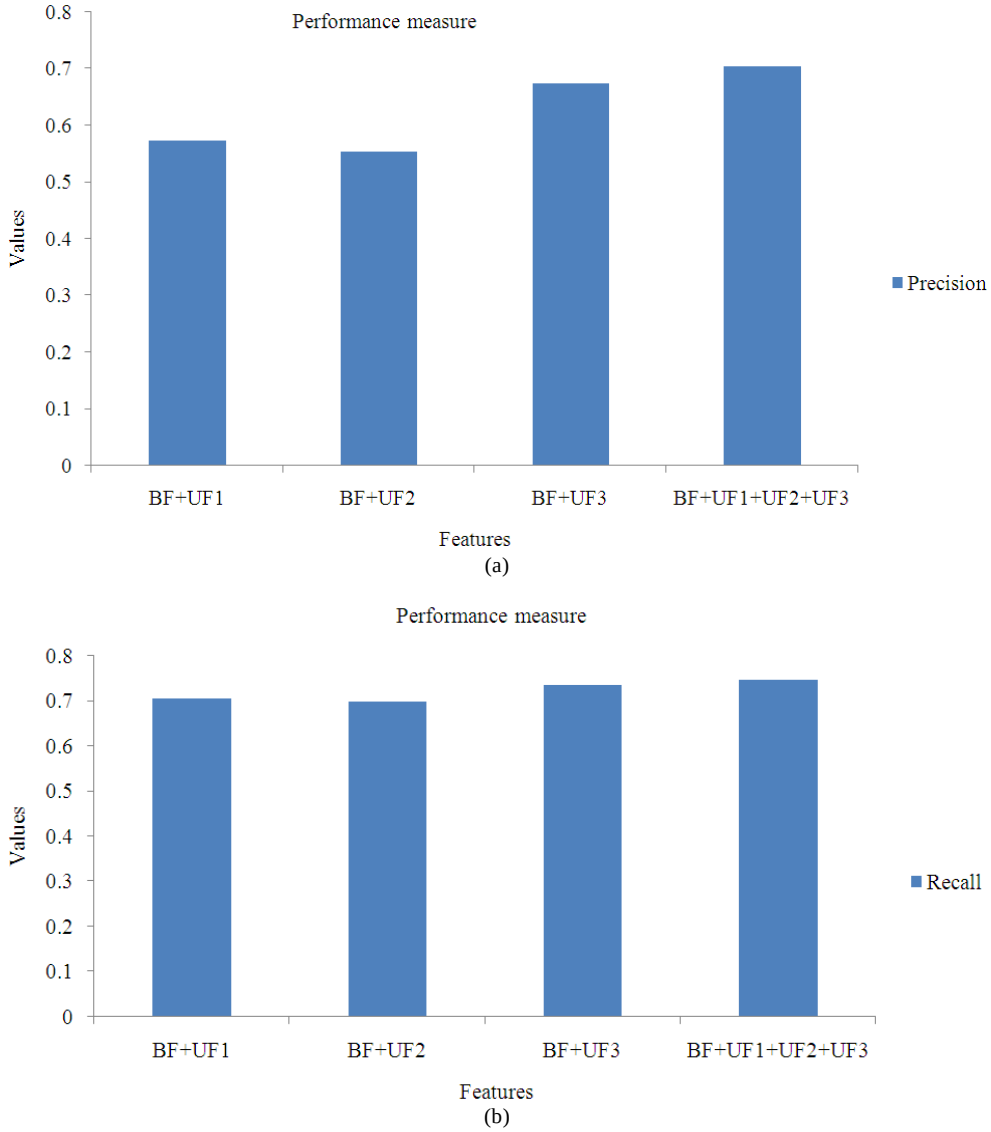


Fig. 5: (a) Update summary-precision (b) Update summary-recall

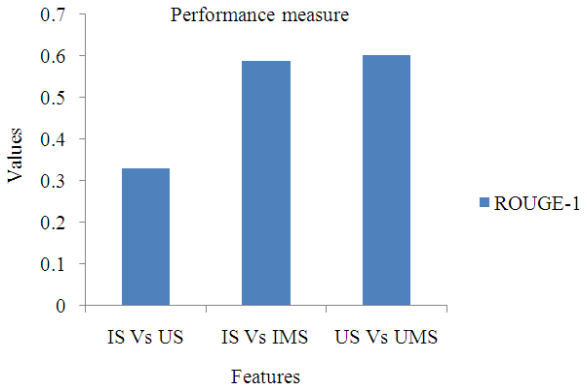


Fig. 6: ROUGE-1 measure

ROUGE-1 measure: To evaluate automatic summary, Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is used. ROUGE measures the quality of a summary by counting the overlapping units such as the n-gram, word sequences and word pairs between the generated summary and the manual summary. We use ROUGE-1 as the evaluation metric Eq. 11:

$$\text{ROUGE\_1 Score} = \frac{X}{Y} \quad (11)$$

where, X is a count of unigrams that occur in the machine and manual summary and Y is a count of unigrams. The following Fig. 6 compares ROUGE-1 Score of Initial Summary(IS) with Update Summary(US), Initial Summary with Initial Manual Summary (IMS), Update Summary with Update Manual Summary(UMS). The Initial Manual Summary and Update Manual Summary are generated manually by us. The result shows that the overlap between Initial Summary and Update Summary is low.

CONCLUSION

The proposed system generates initial and update summary from multiple documents based on annotating the sentences and relevant sentences are selected by utilizing Wikipedia which is used to get concepts and by applying different combinations of features. Relevancy is improved by adopting TSCF - ISF measures. The update summary generated by applying the proposed novel sentence similarity measure is compared with a manual summary as well as with its initial summary and the result shows that the proposed system summary is proficient.

REFERENCES

Barrera, A. and R. Verma, 2010. A ranking-based approach for multiple-document information extraction. University of Houston.

Binwahlan, M.S., N. Salim and L. Suanmali, 2009. Fuzzy swarm based text summarization. J. Comput. Sci., 5: 338-346. DOI: 10.3844/jcssp.2009.338.346

Binwahlan, M.S., N. Salim and L. Suanmali, 2010. Fuzzy swarm diversity hybrid model for text summarization. Inform. Process. Manage., 46: 571-588. DOI: 10.1016/j.ipm.2010.03.004

Jin, F., M.L. Huang and X.Y. Zhu, 2011. Guided structure-aware review summarization. J. Comput. Sci. Technol., 26: 676-684. DOI: 10.1007/s11390-011-1167-y

Kogilavani, A. and P. Balasubramanie, 2011a. Semantic summary generation from multiple documents using feature specific sentence ranking strategy. Elixir J. Comput. Sci. Eng., 40: 5372-5375.

Kogilavani, A. and P. Balasubramanie, 2011b. Multi-document summarization using genetic algorithm-based sentence extraction. Int. J. Comput. Appli. Technol., 40: 246-253. DOI: 10.1504/IJCAT.2011.041653

Kumar, Y.J. and N. Salim, 2011. Automatic multi document summarization approaches. J. Comput. Sci., 8: 133-140. DOI: 10.3844/jcssp.2012.133.140

Long, C., M.L. Huang, X.Y. Zhu and M. Li, 2010. A new approach for multi-document update summarization. J. Comput. Sci. Technol., 25: 739-749. DOI: 10.1007/s11390-010-9361-x

Owczarzak, K. and H.T. Dang, 2011. Who wrote what where: Analyzing the content of human and automatic summaries. Proceedings of the Workshop on Auomatic Summarization for Different Genres, Media and Languages, Jun. 23-23, Oregon, Portland, pp: 25-32.

Varma, V., P. Bysani, K. Reddy, V.B. Reddy and S. Kovelamudi *et al*., 2010. IIIT Hyderabad in guided summarization and knowledge base population. International Institute of Information Technology.