RESEARCH ARTICLE OPEN ACCESS

A Model for Automatic Abstractive Multidocument Summarization

Hadia Abbas Mohammed elsied [1], Naomie Salem [2], Atif Khan [3]
Rihab eltayb Ahmed [4], Yogan Jaya Kumar [5]
Sudan University of Science &Technology [1]
University Teknologi Malaysia [2] & [3], 81310 Skudai, Johor, Malaysia
Univesity of Tabuk [4]
University Teknikal Malaysia Melaka [5], 76100 Melaka, Malaysia

ABSTRACT

We propose a model for multi document abstractive summarization based on Semantic Role Labeling (SRL (in which the content of the summary is not from the source document but from the semantic representation of the source document.. In this model we employ SRL to source document to represent the text source semantically as Predicate Argument Structures (PAS's). Content selection for summary is made by combining the PAS's based on the Cross document Structure theory(CST) relations that each PAS has with other PAS's, then according to number of relation types that each PAS holds we give a score to each PAS ,then we combine the PAS's according to rules related to CST so as to reduce redundancy next the PAS's were ranked using document no and the sentence position No in that document , lastly the selected higher scored PAS's are o final summary. The Experiment for this study supposed to be carried out using DUC 2002 , the standard corpus for text summarization

Keywords:- Abstractive summary, Semantic Role Labeling(SRL), Cross document Structure Theory(CST).

I. INTRODUCTION

As Information overload One of the difficulties that people phased such as many documents share similar topics. On one hand, the similar information which are shown by several different documents, become very tedious for the end users, as they have to read the same information repeatedly . On the other hand, such redundancy can be used to establish significant effort on an applications based on these information as the automatic multi document summarization. Therefore the summaries which synthesize such redundancy across multiple documents would be more beneficial and useful for users as they reduce their time for finding important information across multiple documents. The automatic multi document summarization is deemed one of the major fields of natural language processing (NLP) and it attracts many researchers in recent years[1]. In this study we will automatically synthesize similar information across multiple documents and language generator to produce automatic abstractive multi document summary.

Text Summarization is the process of producing a shorter presentation of original content which covers non-redundant and salient information extracted from a single or multiple documents. The task of summarization is achieved by humans after reading and understanding the document content then selecting the most important points and paraphrasing them into a concise version. Since this procedure takes a long time for one document to be summarized by humans , automatic summarization is coming to the picture. Automatic summarization systems condense documents by extracting the most relevant facts using computer machines. Text summarization methods can be divided into two main approaches extractive approach and abstractive approach. Majority of the studies concentrate on Extractive approach which uses techniques of sentence extraction [2] ,statistical analysis[3][4] , and machine learning techniques[5]. Generally the extractive approach extracts important sections of the text and reproduce them verbatim [6]. In contrast, the abstractive approach aims to produce the important ideas in the text using new phrases or new compression version of the original source [7][8] .Therefore Abstractive summarization is a challenging area as it require deeper analysis of text. The target of abstractive summarization is to improve the generated summary and reduce its redundancy moreover to increase the coherency [9].

The rest of this paper is organized as follows: Section 2 presents related works. Section 3 outlines the proposed model of abstractive multi document summarization the feature extraction will be introduced in section 4 followed by cross document structure theory identification in section 5, PAS scoring and combining will be shown in section 6 and 7 consequently , in section 8 we will go through ranking PAS's , the generation of the abstractive summary will be presented in section 9 , experiment setting and results in section 10 and 11 , and finally end with the conclusion in section 12.

II. RELATED WORK

Abstractive summarization techniques are generally classified into two categories: linguistic based approach and semantic based approaches The former based approaches goes to analyze the text syntactically and uses the syntactic parser to identify the verbs and nouns which are used as a representation of the text , later this representation are used to generate the abstractive summary, where the latter the semantic based methods uses semantic representation of documents to feed in to (NLG) system. Example of semantic representation used in previous studies are ontology based and template based representation [10].

Previous linguistic based approaches which are proposed by [1][8] rely mainly on syntactic representation of source document, these approaches uses syntactic parser for such syntactic representation. The disadvantage of these approaches is the lack of semantic representation of source text .Consequently the semantic approach is considered more appropriate as it based on semantic representation for the source text.

On the other side , a few studies on semantic based methods have also been introduced for abstractive summarization .In the subsequent lines we discussed briefly on them . A multi-document summarization system, name as GISTEXTER, presented in [10] proposed template based method to produce abstractive summary from multiple newswire/newspaper documents depending on the output of the information extraction (IE) system. The template used for topic representation of document. The major limitation found in this approach was that linguistic patterns and extraction rules for template slots were manually created by humans, which is considered time consuming. in addition to that, this method could not handle the information about similarities and differences across multi documents.

Another approach that is a fuzzy ontology based approach [11] was introduced for Chinese news summarization to model uncertain information but in spite of its better description of domain knowledge, the approach has several side effects. First, the time consumed by defining domain ontology and Chinese dictionary by a domain expert. Secondly, Its limitation to Chinese news, and might not be applicable to English news.

The abstractive approach presented by [11] shows a new semantic graph called Rich Semantic Graph (RSG). RSG is an ontology-based representation developed to be used as an intermediate representation for Natural Language Processing (NLP) applications. The new approach consists of three phases: creating a rich semantic graph for the source document, reducing the generated rich semantic graph to more abstracted graph, and finally generate the abstractive summary from the abstracted rich semantic graph. The verbs and nouns of the input document are represented as graph nodes along with edges corresponding to semantic and topological relations between them .The limitation of this approach was that it also relies on manually built ontology, which is time consuming.

The abstractive generated summary which proposed by [7] shows that full abstraction can be achieved in the context of guided summarization. the proposed methodology relies on Information Extraction and Natural Language Generation, uses a rule based information extraction module, content selection heuristics and one or more patterns for sentence generation. The limitation of the methodology was that both information extraction (IE) rules and generation patterns were written manually , which was again time consuming.

The limitation of most semantic based approaches for abstractive summarization is that they almost human dependent that is rely on human expert to construct the domain ontology and rules and this is considered a major drawback for an automatic summarization system. in this study we are looking for a treatment to all these mentioned drawbacks by employing semantic role labelling technique SRL to the source document and extract predicate argument structure as semantic representation automatically.

SRL is previously used by [12] to produce a good extractive summarization , the researcher first employ the SRL and extract the roles from each sentence then for each pair of sentences calculates the semantic similarity based on WorldNet for sentences arguments. Also SRL is used by [13] to produce a framework for multi-document abstractive summarization . First they employed SRL to extract predicate argument structure from each sentence in the document collection as a semantic representation of the text .. In contrast our proposed study uses SRL for abstractive summarization in which we extract predicate argument structure (PAS) from each sentence and for each verb/predicate in the sentence .Moreover we extract the features from each PAS which is not considered statistical one's but more semantic ones such as PAS to PAS semantic similarity, NP to NP semantic similarity ,VP to VP semantic similarity , synonym overlap between PAS's and PAS length .The semantic similarity here we used is based on Jiang similarity which work on concepts of wordnet taxonomy . [13] construct a semantic similarity matrix for scoring each pair of predicate argument structure based on semantic features and text features , in contrast in our study we identify cross document structure theory relationship (CST) [14] between each pair of PAS's using the semantic features we mentioned earlier and further combine the PAS's based on CST relations they hold and next we score the PAS's using two combined measures , first scoring the PAS's depending on the number and type of CST relations that each PAS holds with other PAS's in the document collection and the semantic similarity score for each PAS measure.

III. A MODEL FOR AUTOMATIC ABSTRACTIVE MULTI DOCUMENT SUMMARIZATION

3.1 Overview of Approach

Figure1 depicts our proposed the frame work in which we have a collection of documents supposed to be summarized . The first step is to split each sentence in the document collection that is each sentence is preceded by its corresponding document number and sentence position number. The next step is to employ SENNA semantic role labeler [15] so as to extract the Predicate Argument Structure (PAS) from each sentence in the document collection. In next step we need to identify the Cross Document Structure(CST)[16] between each pair of PAS's .After that we need to combine these PAS's based on their identified CST in previous step. Moreover we Score the PAS's according to the number of CST that each PAS contains therefore we select small number of PAS's with respect to the summary ratio , Finally the selected PAS's will be ordered to form the final summary.

3.2 Semantic Role Labeling (SRL)

Semantic role labeling (SRL) is a process to identify and label arguments in a text. SRL can be extended for the events characterization task that answer simple questions such as “who” did “what” to “whom”, “where”, “when”, and “how”. The main task of SRL is to show what specific relations hold among a predicate with respect to its associated participants . SRL aims to identify the constituents of a sentence, with their roles such as Agent, Patient, Instrument etc., and the adjunctive arguments of the predicate such as Locative, Temporal, with respect to the sentence predicates[18]. This type of role labeling thus produce a first level semantic representation of the text that indicates the basic event properties and relations among relevant entities that are expressed in the sentence [19].

in this study we employ SRL to extract the Predicate Argument Structure (PAS) to be as a representation for our dataset, for the SRL we use SENNA toolkit . SENNA is a software distributed under a non-commercial license, which produces a host of Natural Language Processing (NLP) predictions: semantic role labeling (SRL) ,part-of-speech (POS) tags, chunking (CHK) and name entity recognition (NER). As a preprocess for our dataset we decompose the document collection to sentences , each sentence is preceded by its document number and sentence position number , next we employ the SRL to parse each sentence and label the semantic phrases /words in each sentence properly , we referred to these phrases as semantic arguments . Semantic arguments are accumulated in tow groups :

core arguments (Arg) and adjunctive arguments (ArgM) as illustrated in Table 1. In this study, we consider A0 for subject, A1 for object, A2 for indirect object as core arguments, and ArgM-LOC for location, ArgM-TMP for time as adjunctive arguments , V for predicate (Verb). We put into account all the complete predicates associated with the single sentence structure so as to avoid loss of important terms/words that participate to the meaning of a sentence and its predicates. We suppose that predicates are complete if they have at least two semantic arguments. The predicate argument structure which is extracted used as semantic representation for each sentence in the document collection. We represent the sentence which contains one predicate by simple predicate argument structure where the sentence which contains more than one predicate will be represented by composite predicate argument structure that is the number of predicates in a sentence is equal to the number of extracted predicate argument structure extracted from the same sentence .

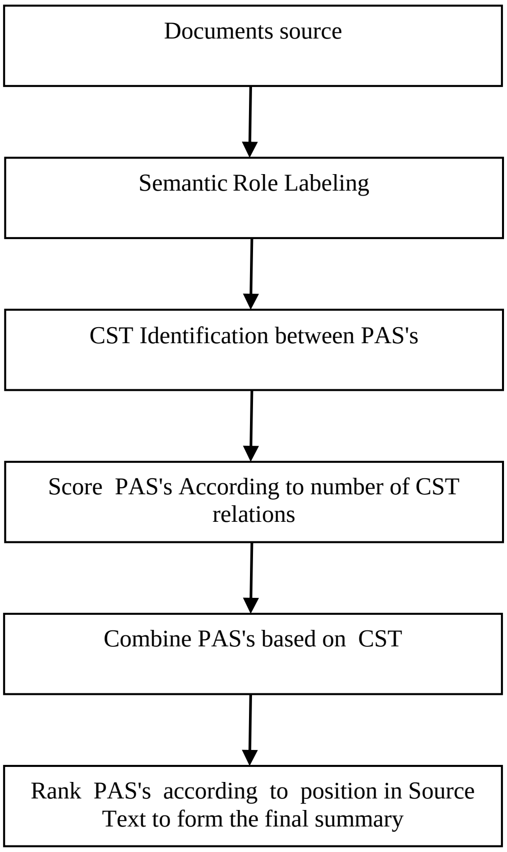

Documents source
Semantic Role Labeling
CST Identification between PAS's
Score PAS's According to number of CST relations
Combine PAS's based on CST
Rank PAS's according to position in Source Text to form the final summary

Figure 1. The proposed Frame Work

Table 1. Representation of Core Arguments and Adjunctive Arguments

Core Arguments		Adjunctive Arguments	
V	verb	ArgM-ADV	adverbial modification
A0	subject	ArgM-DIR	direction
A1	object	ArgM-DIS	discourse marker
A2	Indirect object	ArgM-EXT	extent marker
A3	Start point	ArgM-LOC	location
A4	End point	ArgM-MNR	manner
A5	Direction	ArgM-MOD	general modification
		ArgM-NEG	negation
		ArgM-PRD	secondary predicate
		ArgM- Table 1. PRP	purpose
		ArgM-REC	reciprocal
		ArgM-TMP	temporal marker

Table 2. The algorithm of Multi document Abstractive summarization based on CST Relation

Step	Main process	Process detail
1	Input document collection D	Take the document set as input , $\mathbf{D = \{D_1, D_2, \ldots D_{n1}\}}$
2	Employ SRL for each document	2.1 Employ the SRL for each document using SENNA, this will result as SRL output file. 2.2 Extract the PAS's from each sentence in the SRL output file .
3	CST Identification between each pair of PAS's	3.1 Use the PAS's extracted from annotated CSTBank data set by employing step (2). 3.2 Train the CBR classifier with the CSTBank PAS's and test it using PAS's extracted from DUC data set. 3.3 obtain the CST relations for each PAS in the collection
4	Score PAS's according to CST relations	4.1 Give a score to each PAS , the score will be higher with respect to number of CST relations that each PAS holds, refer to Eq11. 4.2 select top scored PAS's.
5	Combine the PAS's based on their CST relations they hold	5.1 Refer to the rules shown in figure 5 in section 5
6	Order PAS's according to position in Source Text	6.1 each PAS is preceded with doc number and sentence position number in that doc 6.2 order PAS's according to (6.1)

IV. FEATURE EXTRACTION

Every PAS pair will be represented by its feature vector. The features are selected to adapt the related task to the problem of determining rhetorical status from texts. In this paper, five features which compound of (deeper syntactic-level features) are unique to our cross-document relationship types are selected to represent each PAS pair[20]. The features include: Synonym Overlap (SO), Noun-Phrase(NP) and Verb-Phrase(VP) Similarity from each PAS pair based on Jiang Similarity , PAS to PAS similarity based on Jiang Similarity and PAS Length. Below we provide the feature description for each of the mentioned features:

4.1 Synonyms Overlap in PAS (SOP)

This feature represents the measure based on the number of overlapping words or synonyms of words based on wordNet between the two PAS's [21].

$$SOP = \frac{\#\ CW\ orSN(P1,P2)}{\#w\ or\ SN(P1) + \#w\ or\ SN(P2)} \qquad (1)$$

where CW is the # of common words , SN is Synonym of words and P1,P2 are the PAS's undergo overlapping, w is the # of words.

4.2 Type Length based one length of PAS TL(PAS)

This feature is calculated as a ratio of the number of words in PAS over the number of words in the longest PAS in the document[22].

$$L(PAS) = \frac{\#W\ occuring\ \ in\ the\ PAS}{\#\ W\ occuring\ in\ the\ longest\ PAS} \quad (2)$$

TL(PAS1) =1
if L (PAS1>L (PAS2),
= -1 if L (PAS1< L (PAS2),
= 0 if L (PAS1)=(L (PAS2).

where L is the Length of PAS and TL is Type length of PAS which its value based on length of PAS.

4.3 Noun-Phrase(NP) Semantic Similarity

This feature determines semantic similarity between Noun-Phrases in each pair of PAS using Jiang semantic similarity measure. The head tokens of NP in PAS1 and PAS2 are extracted and considered for semantic similarity [23][24].

$$Sem_{NP}(p_i,p_j) = Sem(NP_i\ ,NP_j\ ) \qquad (3)$$

where NP is the for Noun-Phrase , i ,j discriminate Noun-Phrase in PAS1 and Noun-Phrase in PAS2 .

4.4 Verb-Phrase(VP) Semantic Similarity

This feature determines semantic similarity between Verb-Phrase similarity in each pair of PAS Jiang semantic similarity measure. We extract the head token of VP of PAS1 and the head token of VP of PAS2 and then calculate the similarity between them [23][24].

$$Sem_{VP}(p_i,p_j) = Sem(VP_i\ ,VP_j\ ) \qquad (4)$$

where VP is the for Verb-Phrase , i ,j discriminate Verb-Phrase in PAS1 and Verb-Phrase in PAS2.

4.5 PAS to PAS Semantic Similarity

This feature computes the semantic similarity between pair of predicate argument structures. To compute the similarity between two PAS's (P_i, P_j)we calculate similarity for each argument in PAS P_i with its corresponding one in PAS P_j (if no corresponding argument the similarity will be zero) as shown below :

$$sem_{arg}(P_i, P_j) = sem(A0_i, A0_j) + sem(A1_i, A1_j) + sem(A2_i, A2_j) \quad (5)$$

$$sem_V(P_i, P_j) = (sem(V_i , V_j)) \quad (6)$$

$$(P_i, P_j) = (sem(Tmp_i , Tmp_j)) \quad (7)$$

$$sem_{loc}(P_i, P_j) = (sem(Loc_i , Loc_j)) \quad (8)$$

We combine (5), (6), (7), and (8) to give (9).

$$sem (P_i, P_j) = sem_{arg}(P_i, P_j) + sem_V(P_i, P_j) + sem_{tmp}(P_i, P_j) + sem_{loc}(P_i, P_j) (9)$$

where A0,A1,A2,V,Tmp,Loc as shown in Table 1.

V. CROSS DOCUMENTS STRUCTURE RELATIONSHIP THEORY (CST)

In this paper , we will investigate the utility of multi document analysis. Discourse analysis in texts currently become very dominant, specifically when it involves multiple texts i.e. documents news. the Information across topically related documents can often be connected. The idea of cross-document relation identification is to study the existence of inter-document relationships between texts. The cross-document relations are based on the Cross-document Structure Theory (CST) model which was introduced by Radev [23], who explores that documents which are related to the same topic will contain semantically related textual units. Moreover he analyzed and investigated the relationships that might exist between sentences across the related documents.

The motivation of the use of CST relationships in this study is lying for two reasons , first is that our study is conducted upon multi document abstractive summarization which is experimented using the DUC 2002 dataset which incorporated from set of related documents where we can find CST relations between them, second we believe that the important information expressed in a sentence of a document is also expressed in the sentences of many related document besides the other sentences within the same document , for these two mentioned reasons we include the number of CST Relations that each PAS holds to contribute in the final score of a PAS .

5.1 Identification Of CST Relations between PAS

We propose the use of CST for multi document abstractive summarization to include and rank the documents sentences based on their CST relations for summary generation. Previous works on multi document summarization based on CST relations are employed for extractive summaries in which they regarded the CST types separately. Moreover they use plain text as data set , where in our work which is mainly for abstractive multi document summarization we concentrate on the PAS's as representation of plain texts. We need to identify the CST relations among each pair of PAS's.

Earlier previous works based on CST, regarded the CST types separately [25], where we in this study investigate the combination of some types of CST to give a new CST because of their similar characteristics.

According to the definition by CST, some of the relationship presents similar surface characteristics. Except for different version of event description, relations such as Paraphrase, Modality and Attribution share similar characteristic of information content with Identity[26].Table 3 shows Part of the CST Relations , and further details can be found in [29].

Table 3. CST Relations

1	Identity	12	Indirect speech
2	Equivalence (Paraphrase)	13	Elaboration (Refinement)
3	Translation	14	Fulfillment
4	Subsumption	15	Description
5	Contradiction	16	Reader Profile
6	Historical Background	17	Change of per-spective
7	Citation	18	Overlap (partial equivalence)
8	Modality		
9	Attribution		
10	Summary		
11	Follow-up		

In this study we propose five types of CST Relations which are Identity , Subsumption , Overlap, Description and No Relations .three types of them resulted as a combination of other types according to their similar surface characteristics as highlighted in table. 4.

For the identification of these CST relations , we develop the CBR classifier by i.e Case Base Reasoning (CBR)[21], we extract relevant features from each PAS pairs.

Figure2 demonstrated the overview of the proposed method for cross-document relation identification. In order to identify the relationship type between PAS pairs, each PAS pair will subjected feature extraction which supplied to CBR classifier. The specifics descriptions of these processes are given in the following sections.

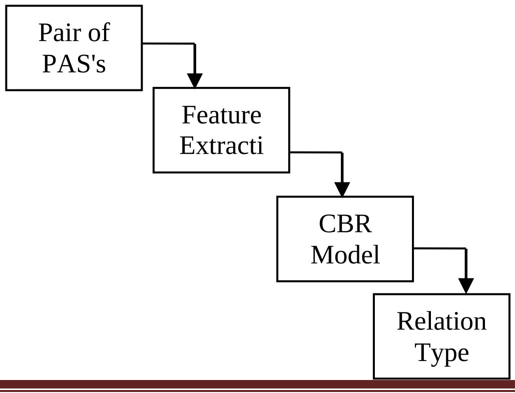


Figure2: Proposed method for CST Identification

Table 4. The proposed CST relations

Relations by CST	Proposed Relations	proposed CST NO	Definition of Proposed Relation
Identity 1, Paraphrase 2, Modality 8, Attribution 9	Identity	1	Two text spans have the same information content
Subsumption 4, Elaboration 13	Subsumption	4	S_1 contains all information in S_2, plus
Overlap 18	Overlap	1 8	S_1 provides facts X
Description 15, Historical Background 6	Description	15	S_1 gives historical context or describes an
-	No Relations	0	No relation exits between S_1 and S_2.

5.2 Case Based Reasoning Approach

Case Based Reasoning (CBR) is a family of artificial intelligence techniques, based on human problem solving paradigm [27]. CBR is different from other AI approaches , while not relying on general knowledge of problem CBR is able to utilize its knowledge base domain of previously solved problem and concrete problem situations (cases). Anew problem will be solved by benefited from previous similar cases which called "Reuse" .Also another characteristics for CBR that it lies on its ability to incremental , saving new solutions and this is called "Retain" which will widen the chance to solve new problems

5.2.1The CBR cycle

For example, when a new case is input into the CBR cycle, the following steps will be taken to solve it:

1. Retrieve – the most similar cases from the case base;
2. Reuse– the solutions from the retrieved cases;
3. Revise – the solution for the new case if necessary
4. Retain – adapt revised new cases into the case base.

A new problem is solved by retrieving one or more previously experienced cases, reusing the case in one way or another, revising the solution based on reusing a previous case, and retaining the new experience by incorporating it into the existing knowledge base (case-base) [28]. The four processes are illustrated in Figure 3

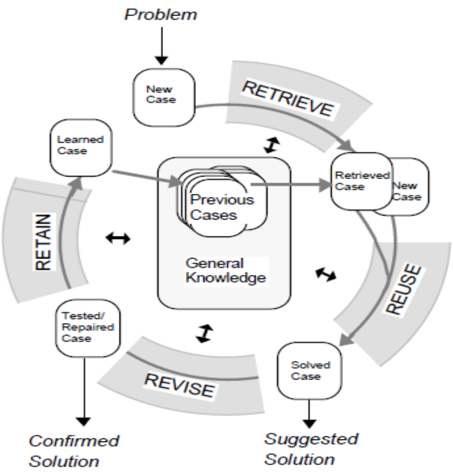
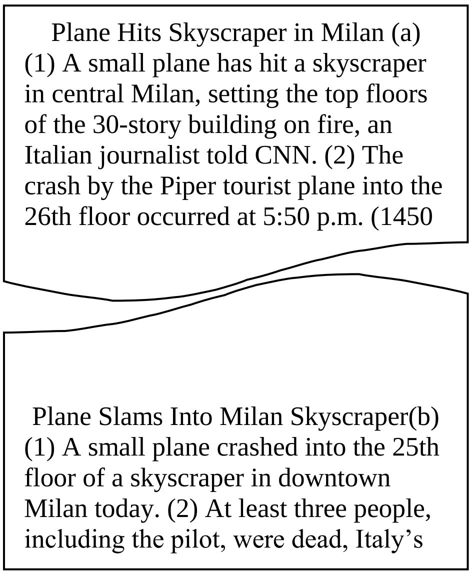


Figure3: CBR life cycle [28]

5.2.2 The uses of CBR in this study

In this study we use CSTBank dataset which annotated by CST relations .The CST data set compose of examples of sentence pairs annotated with cross document relations for example see Figure 4 , we observe that in the first document (a) Sentences 2 and in the second document (b) sentence 2 contradict each other (25th floor vs. 26th floor)[29]. For our study we will employ SRL for the dataset and extract PAS's along with the features mentioned in section 3.Therefore we enrich the CBR Knowledge base with the pair of PAS's features with respect to their annotated CST type. we train the CBR using the enriched PAS and then we use the DUC 2002 PAS's as testing ,accordingly the CBR output which represented as CST type will be considered as the relation type identification.



Plane Hits Skyscraper in Milan (a)
(1) A small plane has hit a skyscraper in central Milan, setting the top floors of the 30-story building on fire, an Italian journalist told CNN. (2) The crash by the Piper tourist plane into the 26th floor occurred at 5:50 p.m. (1450

Plane Slams Into Milan Skyscraper(b)
(1) A small plane crashed into the 25th floor of a skyscraper in downtown Milan today. (2) At least three people, including the pilot, were dead, Italy’s

Figure 4 : Snapshot of CSTBank data set examples

VI. PAS'S SCORING

Normally as In CST, researchers follow a common approach that is to select the sentences with high number of relations . Here in this study, Since not all CST relations contributing in equal way in the summary , we suggest to assign the maximum similarity for the PAS with other PAS's in the document in addition to the Number of CST relations which hold by the same PAS with other PAS's in the document set , and this is considered as a final score for each PAS In order to get the best results. to achieve that we will use equation Eq9, Eq10 and Eq11.

$$PAS\_CST\_Score = \sum_{i=1}^{n} \#\,\text{Relation} \qquad (10)$$

where n is the number of CST relations that PAS holds.

Since not all CST relations contribute equally in the summary we need to refine (Eq11) by adding a sort of fairness regarding the distributions of CST relations for each PAS with other PAS's in the document set ,to do this we look forward to divide the number of CST relations that each PAS hold by the total number of the PAS's in document set as shown in Eq12.

$$PAS\_CST\_Score\_Ratio(p) = \frac{\sum_{i=1}^{n} \#\,\text{Relation}}{\sum_{j=1}^{N} \#\,\text{PAS} \;-1} \qquad (11)$$

where $\sum_{i=1}^{n} \#\,\text{Relation}$ is the total number of CST relations that one PAS hold with all other PAS's in the document set , this total is divided by the total number of PAS's in the document set subtracted from it the current PAS which is indicated by $\sum_{j=1}^{N} \#\,\text{PAS} - 1$.

$$FinalPAS_{score} = PAS\_CST\_Score\_Ratio(p) + max(sim(P)) \qquad (12)$$

where sim (P) calculated by using (Eq 9) then we select the maximum similarity that the current PAS have with other PAS's in the document set in addition to $PAS\_CST\_Score\_Ratio(p)$ which calculated using (Eq11).

VII. COMBINE PAS'S BASED ON CST

In this phase we suggest to combine each pair of PAS's that hold specific CST type so as to reduce redundency using the proposed rules illustrated in figure 5.

If CST is one of the following Do the decision:
Case CST between P1 And P2 is Identity : select P1
Case CST between P1 And P2 is Subsumption : select p1
Case CST between P1 And P2 is Equivelance : select P1.
Case CST between P1 And P2 is Historical : select P2.
Case CST between P1 And P2 is Contradiction :do not select
Case CST between P1 And P2 is no relation : select Both

Figure5.CST Relations Combination Rules

VIII. Rank PAS's according to position in Source Text

The ordering of the PAS's will be according to the document number and the sentence position number which are previously attached to each PAS, now they will be used to accomplish this task.

IX. ABSTRACTIVE SUMMARY GENERATION

Generation of abstractive summary will be accomplished by combining the ordered PAS's in previous phase in a way that all PAS's which share the same subject in other words having same A0 will be combined by removing the subject from all PAS's except the first one and connecting them using "and" operator, if the PAS's are more than two other wise using comma to separated them and "and" connector to connect the last PAS . We combine each pair of PAS’s according to the rules suggested in figure 5 , then we select the best (highest (PAS To PAS) semantic similarity) as we can find that each PAS can have relation with many other PAS’s , the overall summary is 20% ratio from all PAS’s , therefore we select the best 20% of highest PAS to PAS semantic similarity.

X. EXPERIMENT SETTING

In this experiment we want to study whether the combination of PAS according to CST relations has impact on summarization. At first we perform preprocessing on the sets of documents. This step involve sentence splitting , tokenization , removal of stop words and word stemming. Once the document are preprocessed , we apply semantic role labeling (SRL) technique to extract Predicate Argument Structure (PAS’s) from document sentences. Next we conduct a comparison between each PAS and all other PAS’s in the document to find out the CST (Cross document relation Identification Structure Theory), to accomplish this work first we extract five features from each pair of PAS’s such as SO (Synonym Overlap), NP (Noun Phrase) similarity, VP(verb phrase) similarity , PAS to PAS similarity and PAS length . These features are extracted and calculated by using the equations mentioned earlier in section 4. Next we use a CBR classifier to identify the CST relation between each pair of PAS’s in the document , and then a combination between PAS’s is given to form the final summary .This combination is carried out according to rules suggested by the researcher. We employ three pyramid evaluation measures, mean coverage score (Recall), precision , and F_measure. For evaluation of proposed model for automatic abstractive multi document summarization , this metric evaluates the quality of peer summary (System produced summary) by comparing it with human model summaries and other benchmark summarization system in the context of DUC 2002 multidocument abstractive and extractive summarization shared tasks.

XI. EXPERIMENT RESULTS

The proposed approach is evaluated in the context of multi-document abstractive summarization task, using news articles/data sets provided by the Document Understanding Evaluations 2002. For each data set, our approach generates a summary with 20% compression rate, the task tackled by other systems participating in multi-document abstractive summarization task. To compare the performance of our proposed approach (we call it AS-SRL-CST), we setup four comparison models, which are as follows: AS[7] refers to the recent abstractive approach for multi-document summarization, Best automatic summarization system (Best) in DUC 2002, AS-SRL[13] refers to semantic approach for multi-document abstractive summarization using semantic role labeling in DUC 2002, and the average of human model summaries (Models). For comparative evaluation, Table 5 shows the mean coverage score (recall), average precision and average F-measure obtained on DUC 2002 dataset for the proposed approach (AS-SRL-CST), the Best system, AS-SRL in DUC 2002, and the average of human model summaries (Models). Figure 6 visualizes the summarization results obtained with the proposed approach and other comparison models.

Table 6:Comparison of multi-document abstractive summarization results in DUC 2002 based on mean coverage score, average precision, and average F-measure.

System	**Mean Coverage Score**	**AVG-Precision**	**AVG-F-Measure**
Models	0.6910	0.8528	0.7634
AS-SRL	0.4431	0.60	0.5153
AS(Genest and Lapalme,2011)	0.4378	0.59	0.50
Best	0.2783	0.7452	0.4053
AS-SRL-CST	**0.5457**	**0.4378**	**0.4818**

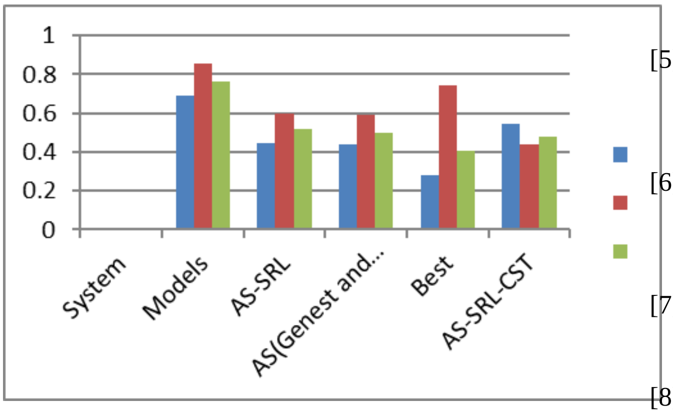


Figure 6: Comparison of summarization results based on mean coverage score, average precision and average F-measure

XII. DISCUSSION

It could be observed from the results given in Table 6, that on mean coverage score the proposed approach (AS-SRL-CST) yields better summarization results than other comparison summarization models; and appeared less better in AVG-Precision and average F-Measure,) , but better than Best.

The drop in precision measure in our proposed approach might be due to the use of non-optimized features for selection of PASs for summary generation.

The experimental finding supports the claim that automatically identified semantic representation extracted from document text using semantic role labeling facilitates the semantic analysis of documents, and thus leads to better summarization results.

XIII. CONCLUSION AND FUTURE WORK

This study demonstrated the work with abstractive multi document summarization which based on the SRL and the extraction of PAS's , more over we identify the relation ship between the extracted PAS's in the form of predefined CST relations which is discovered between each pair of PAS's, in addition to that we give a score to each PAS using equation s mentioned in section 5. We combine the PAS's beneficiaries from the CST between each pair and for that we follow the rules illustrated in figure5.

REFERENCES

[1] R. Barzilay and K. R. Mckeown, “Sentence Fusion for Multidocument News Summarization,” no. September 2003, 2005.

[2] J. Kupiec, J. Pedersen, and F. Chen, “A trainable document summarizer,” … 18th Annu. Int. ACM …, 1995.

[3] M. Chandra, V. Gupta, and S. Paul, “A statistical approach for automatic text summarization by extraction,” Commun. Syst. …, pp. 268–271, Jun. 2011.

[4] K. K. and D. Marcu, “Statistics-Based Summarization -step one : sentence Compression,” Am. Assoc. Artif. Intell., 2000.

[5] J. L. Neto, A. A. Freitas, and C. A. A. Kaestner, “Automatic text summarization using a machine learning approach,” Discovery, vol. 2057, no. i, pp. 205–215, 2003.

[6] V. Gupta and G. S. Lehal, “A Survey of Text Summarization Extractive Techniques,” J. Emerg. Technol. Web Intell., vol. 2, no. 3, pp. 258–268, Aug. 2010.

[7] P. Genest and G. Lapalme, “Fully Abstractive Approach to Guided Summarization,” no. July, pp. 354–358, 2012.

[8] P. Genest and G. Lapalme, “Framework for abstractive summarization using text-to-text generation,” in Workshop on Monolingual Text-To-Text Generation, 2011, no. June, pp. 64–73.

[9] P. Genest and G. Lapalme, “Text generation for abstractive summarization,” Proc. Third Text Anal. Conf. …, 2010.

[10] N. R. Kasture, N. Yargal, N. N. Singh, N. Kulkarni, and V. Mathur, “A Survey on Methods of Abstractive Text Summarization,” no. 6, 2014.

[11] I. Moawad and M. Aref, “Semantic graph reduction approach for abstractive Text Summarization,” Comput. Eng. Syst. (ICCES), …, pp. 132–138, Nov. 2012.

[12] L. Suanmali, N. Salim, and M. S. Binwahlan, “Fuzzy Genetic Semantic Based Text Summarization,” 2011 IEEE Ninth Int. Conf. Dependable, Auton. Secur. Comput., pp. 1184–1191, Dec. 2011.

[13] A. Khan, N. Salim, and Y. Jaya Kumar, “A framework for multi-document abstractive summarization based on semantic role labelling,” Appl. Soft Comput., vol. 30, pp. 737–747, 2015.

[14] D. Radev, J. Otterbacher, and Z. Zhang, “CST Bank: A Corpus for the Study of Cross-document Structural Relationships.,” Lrec, pp. 1783–1786, 2004.

[15] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, “Natural Language Processing (almost) from Scratch,” J. Mach. Learn. Res., vol. 12, no. Aug, pp. 2493–2537, 2011.

[16] D. R. Radev, “Introduction to the Special Issue on Summarization,” pp. 1–11, 2001.

[17] A. Gatt and E. Reiter, “SimpleNLG : A realisation engine for practical applications,” Proc. ENLG ’09 Proc. 12th Eur. Work. Nat. Lang. Gener., no. March, pp. 90–93, 2009.

[18] C. Aksoy, A. Bugdayci, T. Gur, I. Uysal, and F. Can, “Semantic Argument Frequency-Based Multi-Document Summarization,” pp. 470–474, 2009.

[19] L. Màrquez, X. Carreras, K. C. Litkowski, and S. Stevenson, “Semantic Role Labeling: An Introduction to the Special Issue,” Comput. Linguist., vol. 34, pp. 145–159, 2008.

[20] E. G. Maziero, T. Alexandre, and S. Pardo, “Multi-Document Discourse Parsing Using Traditional and Hierarchical Machine Learning,” pp. 1–10, 2011.

[21] Y. J. Kumar, N. Salim, A. Abuobieda, and A. T. Albaham, “Multi document summarization based on news components using fuzzy cross-document relations,” Appl. Soft Comput., vol. 21, pp. 265–279, 2014.

[22] J. Pedersen, X. Palo, A. Research, and P. Alto, “A Trainable Document,” Corpus.

[23] D. Radev, “A common theory of information fusion from multiple text sources. Step one: Cross document structure,” Proc. 1st SIGdial Work. Discourse Dialogue, pp. 74–83, 2000.

[24] J. J. Jiang and D. W. Conrath, “Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy,” no. Rocling X, p. 15, 1997.

[25] M. Lucia, C. Jorge, T. Alexandre, and S. Pardo, “Multi-Document Summarization : Content Selection based on CST Model (Cross-document Structure Theory).”

[26] N. A. H. Zahri, F. Fukumoto, and M. Suguru, “E XPLOITING R HETORICAL R ELATIONS T O M ULTIPLE D OCUMENTS T EXT S UMMARIZATION,” vol. 7, no. 2, pp. 1–22, 2015.

[27] C. Paper, “COMPONENT BASED SUMMARIZATION USING AUTOMATIC IDENTIFICATION OF CROSS- DOCUMENT STRUCTURAL RELATIONSHIP COMPONENT BASED SUMMARIZATION USING AUTOMATIC IDENTIFICATION OF CROSS-DOCUMENT,” no. July, 2015.

[28] A. Aamodt and E. Plaza, “Case-based reasoning: Foundational issues, methodological variations, and system approaches,” AI Commun., vol. 7, no. 1, pp. 39–59, 1994.

[29] D. R. Radev, “Cross-document relationship classification for text summarization,” 2002.