Sudan University of Science and Technology

College of Graduate Studies

A Model for Automatic Abstractive Multidocument Domain-Specific Summarization

نموذج للتلخيص التلقائي للوثائق المتعددة في مجال محدد

A thesis submitted in fulfilment of the requirements for the awarding of the degree of Doctor of Philosophy (Computer Science)

By

Hadia Abbas Mohammed Elsied Ahmed

Supervisor
Prof. Dr.Naomie Binti Salim

March 2019

بِسۡمِ ٱللَّهِ ٱلرَّحۡمَٰنِ ٱلرَّحِيمِ

(" يَرْفَعِ اللَّهُ الَّذِينَ آمَنُوا مِنكُمْ وَالَّذِينَ أُوتُوا الْعِلْمَ دَرَجَاتٍ وَاللَّهُ بِمَا تَعْمَلُونَ خَبِيرٌ)
صدق الله العظيم
سورة المجادلة الآية رقْم 11

DEDICATION

To soul of my mother,

To my father, Husband,

Mydaughter, brothers,

Sisters and all friends

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to all those who provided me the possibility to complete this research. Special gratitude I give to My supervisor prof Dr. NaomieBintiSalim whose contribution to stimulating suggestions and encouragement, helped me to coordinate my research.

Furthermore, I would also like to acknowledge with much appreciation the crucial role of the staff of UTM (University Technology Malaysia) and the staff of SUST whom both gave me the permission to use all required resources to complete the research. A special thanks go to Dr. Atif Khan, Dr.Yogan Jaya Kumar, and Dr.RihabEltaybe, who help me a lot in some specific tasks and gave suggestion about some solutions. Last but not least, many thanks go to the colleagues and friends who always make help and Duaa for me to achieve the objective goal.

ABSTRACT

Documents which are retrieved there on the internet through online search often come with a large amount of text. In the context of news documents, different news sources reporting on the same event usually contain common components that build up the main story of the news. This study aims to provide a new model of multi-document abstractive summarization (SRL-CST) based technique.The study first makes a pre-process to the texts which include sentence splitting, tokenization, stop word elimination and word stemming and then employs the Semantic Role Labeling (SRL) to each sentence and then Predicate Argument Structure (PAS) extracted, which will be the representation of the texts undergo summary.

Since this study involves multiple documents, the research further investigates the automatic identification of cross-document relations from unannotated text documents, where the case-based reasoning (CBR) classification model is proposed. Cross-document relations are used to identify highly relevant sentences to be included in the summary. In the context of CST, the researcher suggests combining each related relation to be in one big relation and this is done based on their similar meaning.

Content selection for the summary is made by combining the PASs based on the Cross document Structure theory(CST) relations that each PAS has with other PASs, then according to number of relation types that each PAS holds a score is given calculated to each PAS ,then we combine the PASs according to rules related to CST suggested by the researcher so as to reduce the redundancy. Next, the PASs was ranked using document No and the sentence position No in that document. lastly, the PASs in the top 20% higher scores are selected to form the final summary. Pyramid evaluation is examined against the study system summary and human model summaries and it could be observed from the results, that on mean coverage score the proposed approach (AS-SRL-CST) yields better summarization results..

المستخلص.

الوثائق والمستندات التي يتم استرجاعها من خلال البحث في الانترنت تأتي بأعداد كبيرة من النصوص . في سياق الوثائق الاخبارية عادة ما تحتوي المصادر الاخبارية المختلفة التي تقدم تقارير عن نفس الحدث على مكونات مشتركة تبنى عليها القصة الرئيسية للاخبار. تهدف هذه الدراسة لتقديم نموذج جديد للتلخيص التجريدي للوثائق المتعدده المبنية على تقنية(SRL-CST) حيث تقوم هذه الدراسة اولا باجراء معالجة مسبقة للنصوص والتي تشمل تقسيم الجملة ,الترميز , ازالة الكلمات التوقفية , الكلمة الجزعية ثم استخدام SRL لكل جملة وأخيراً يتم استخلاص PAS والتي سوف تكون تمثيل للنصوص الخاضعة للتلخيص. وبما ان هذه الدراسة تركز على الوثائق المتعددة فهي ايضاً تسعى لاختبار التحديد التلقائي بين الوثائق من مستندات نصية غير معلومة , لذلك تم اقتراح نموذج التصنيف المنطقي المبني على الحالة(CBR) . أيضاً تم استخدام العلاقات بين الوثائق لتحديد الجمل ذات الملاءمة العالية ليتم تضمينها في الملخص. في اطار تقنية اختيار المحتوى (CST) اقترح الباحث الجمع بين كل العلاقات المتصلة لتبقى في علاقة واحدة شاملة بسبب معانيها المتشابهة.

تتم عملية اختيار المحتوى للملخص من خلال الجمع بين ال(PASs) بناء على علاقات نظرية البنية عبر المستندات (CST) والتي توضح ان كل PAS له علاقة مع ال PASs اخرى ثم تحدد نقاطاً لكل PAS وفقا للقواعد المتعلقة بنظرية اختيار المحتوى التي اقترحها الباحث وذلك للحد من التكرار وبعد ذلك يتم تصنيف الPASs باستخدام رقم الوثيقة او المستند ورقم موقع الجملة في الوثيقة او المستند,اخيراً النقاط ال 20% العليا التي تم اختيارها

هي التي تكون الملخص . تم اختيار التقييم الهرمي مقابل ملخص النظام وملخصات النموذج البشري , حيث يلاحظ من خلال النتايج ان متوسط النقاط في النموذج المقترح تعطي نتائج افضل للتلخيص.

Table Of Contents

LIST OF TABLES

TABLE NO. **TITLE** **PAGE**

LIST OF FIGURES

FIGURE NO. TITLE PAGE

LIST OF APPENDICES

LIST OF ABBREVIATIONS

AS	Abstractive Summarization
CBR	Case Base Reasoning
CHK	Chunking
CST	Cross Document Structure Theory
DF	Document Frequency
DUC	Document Understanding Conferences
IE	Information Extraction
INIT	Information Item
MDAS	Multidocument Abstractive Summarization
NLP	Natural Language Processing
PAS	Predicate Argument Structure
POS	Parts-Of-Speech Tagging
ROUGE	Recall Oriented Understudy for Gisting Evaluation
RS	Rich Semantic Graph
SRL	Semantic Role Labeling
SVM	Support Vector Machine
TS	Text Summarization

CHAPTER I

1.1 Introduction

The world is going very fast towards the information age , this rapid overload of information is referred to the internet delivery , this information can be represented as web pages or text documents, so people succeed to access the online information very easily and since the growth of the internet being day by day, people face the problem of information overload which makes the abstraction summary of retrieved information become very necessary . In the current epoch of Information overload, multi-document summarization is considered an important tool in the field of natural language processing and has won with more concern in recent years (Barzilay&McKeown 2005).

One of the main troubles of the huge information across the internet that many documents share similar topics or events, such redundancy creates a chance for natural languages processing systems.On the other hand the redundancy makes the extraction of information for specific event that repeated in multiple sources be very difficult for the end users as they have to read the information repeatedly across many documents, the redundancy of information can be used to identify the important and significant information for many application such as summarization and question answering . Therefore summaries that fuse information can be useful for end users and as it saves their time for finding the key information (Barzilay&McKeown 2005).

In this research, we propose an approach that will automatically summarize similar event across multi-document news, taking a benefit from cross-document structure theory (CST) to produce a concise abstractive summary.

The needs of automatic summarization are increased nowadays since the information overloaded day by day, therefore, summarization become an integral part of everyday life, for example, the abstract of scientific publications, results retrieved by search engines, the overview of books and newspapers headlines all these examples of summaries. Luhn started the idea of Automatic summarization since 1950 (Luhn 1958), his approach uses term frequencies to measure the sentence relevance i.e. sentences are included in the summary if they contain high frequent terms.

Text summarization methods can be divided into two main approaches: the extractive approach and abstractive approach. The former approach deals with the selection of an important term from the original text and added them to the summary. Here, the text is reduced using the same words mentioned in the original text.The most important content is treated as the most frequent or the most favorably positioned content. The latter approach, the abstractive approach, requires deeper analysis of the text and the ability to generate new sentences, which provide an obvious advantage in improving the focus of a summary and reducing its redundancy(Genest& Lapalme 2011)(Genest& Lapalme 2012). In this study, we will focus on the abstractive approach.

The majority of studies have focused on extractive summarization using a various technique such as sentence extraction(Kupiec et al. 1995) , statistical analysis(Knight &Marcu 2000), discourse structures and other techniques. On other hand, abstractive summarization is a challenging area and is a hope of researchers (Luhn 1958) because it requires deep analysis of the text and has the capability to synthesize novel sentences, which improves the focus of a summary, reduces its redundancy and keeps a good compression rate(Genest& Lapalme 2011). A few semantic-based approaches have been proposed for multi-document abstractive summarization. These approaches employ humanly built domain ontology and template for semantic representation of documents. The obvious issue with these approaches is that they rely on human experts to build domain ontology or design template rules which requires more effort and time, and is a drawback of automatic text summarization. Moreover, these approaches deal with a specific domain and may not be adapted to other domains.

Summarization has become very important due to the daily increasing information on the internet. In order to go through this information in a short time, a reader needs a summarized version out of this information. Since it is difficult for a human to carry out the summary for large texts, automatic summarization is coming to the picture.

There is no standard model for abstractive summarization; therefore abstractive summarization needs more research. In this study, we try to introduce a model for an abstractive summary.

1.2 Problem Background

As the internet is getting fast and fast, the information retrieval also is getting fast, and due to this speed we always receive a huge amount of online text documents when we enquire about specific event, but it will be a very tedious task to read all these retrieved documents which almost contains repeated information therefore automatic text summarization will be a suitable solution and instead of going through all these documents one by one and read repeatedly, a summary which synthesizes common information across many text documents will offer the general conclusion of long texts and this will be useful for the users and save their time for finding the key information in text documents.

Automatic text summarization is to employ the machine to simulate the human work in creating summaries which are considered as a challenging task. An automatic text summarization works by choosing salient sentences from the document text and combining them together, Nowadays people implement two types of text summarization according to a number of documents being summarized, single document and multi-document .summarization.The process of producing a summary from a single document is called a single document summarization. On the other hand, multi-document text summarization aims to help users in managing the great volume of available text documents in digital media, by extracting the most significant common facts or topics across the set of documents.

Multi-document Text Summarization can be classified according to the type of summary: extractive, and abstractive. The extractive summary is the procedure of identifying important sections of the text and producing them verbatim while abstractive summary aims to produce important materials in a new generalized form since it requires deeper analysis of a text. The problem is that in the extractive approach, the produced summary was not guaranteed to be always coherent, which is considered as an important condition for the summary. Therefore, it is more convenient for people to use the abstractive approach to guarantee the coherency as well as consistency although abstractive approach needs more effort than extractive approach.

Abstractive summarization techniques can again be classified into two categories-structured based and semantic-based methods. Structured based approaches determines the most important information through documents by using templates, extraction rules and other structures like tree-based, template based, ontology-based, lead and body phrase and rule-based whereas Semantic-based approach that focus on semantic representation of texts such as identifying the predicate-argument structure of each sentence (Kasture et al. 2014).

Many researchers have tried to generate abstractive summaries using various techniques such as lexical chains which are used by(Barzilay et al. 1999) (Barzilay&Elhadad 1997) where the researchers found the most important concepts statistically rather than through deep semantic meaning which leads to an incoherent summary.

(Liu & Liu 2009) used compression technique, where different compression algorithms were used. The experiments on the corpus showed that abstractive summaries using sentence compression has better ROUGE scores compared to extractive summaries However, the best performance is still quite low, suggesting the need of language generation for abstractive summarization.

(Barzilay&McKeown 2005) proposes a sentence fusion technique. They found that while the output of existing compression algorithm is always a substring of the original sentence, sentence fusion may generate a new sentence which is not a substring of any of the input sentences. This is achieved by arranging fragments of several input sentences into one sentence and this was considered as an advantage over sentence compression.

Moawad and Aref (Moawad&Aref 2012) propose a novel technique that was the Rich Semantic Graph (RSG). The researchers showed that this technique reduces the original text to fifty percent when applied to the case study. The drawback is the evaluation of this technique should be assessed after applying it to more than one case study to produce a proper evaluation.

In this study, 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 (PASs). Content selection for summary is made by combining the PASs based on the Cross document Structure theory(CST) relations that each PAS has with other PASs , then we give a score to each PAS according to the number of relation types that each PAS holds, then the selected 20% higher scored

PASs are ranked to form the final summary. The Experiment for this study is carried out using DUC 2002, the standard corpus for text summarization.

Since this study involves multi-documents, studies related to multi-document analysis is also investigated in this research. Discourse analysis in texts has nowadays become very prominent; especially when it involves multiple texts. One example of such analysis is the study on cross-document relation. The idea of cross-document relation is to investigate the existence of inter-document rhetorical relationships between texts. These rhetorical relations are based on the CST (Cross-document Structure Theory) model(Radev 2000).

Documents which are related to the same topic usually contain semantically-related textual units. For instance, the relation between two textual units (e.g. between two sentences) can be identical, overlapping, descriptive, contradictive and etc. therefore in this study we combine the PASs based on the relation each PAS has with other PASs. Further details on CST relations can be found in Chapter 2.

1.3 Problem Statement

Abstractive summarization is regarded as a significant part of multi-document text summarization it requires deeper analysis of a text. The limitation of all semantic approaches, that they mainly depend on human experts to construct domain ontology and rules and then the semantic representation of source document is built from them and this deemed as a drawback of an automatic summarization system. Therefore, a robust multi-document Abstractive summarization (AS) model must be introduced. Such model is based on the use of Semantic Role Labeling for Predicate Argument Structure (PAS) formation as a representation of source documents, content selection for summary is made by combining the PASs based on the Cross document Structure theory(CST) relations that each PAS has with other PASs, then according to number of relation types that each PAS holds we give a score to each PAS lastly the 20% selected higher scored PASs are ordered to form the final summary. So we want to prove that:

"Predicate Argument Structure (PASs) that created from SRL (Semantic Role Labeling) can be used to generate good abstractive summary "

1.4 Research Question/ Hypothesis/ Philosophy

1.4.1 Research Question

Can Semantic Roles extracted from text be combined using CST relations to generate good abstractive summaries?

1.4.1 .1 Sub-Questions

1- What features of SRL can be used for AS?

2- How can the SRL components be weighted and selected for AS?

3- What are the CST relations that can be used to combine PASs.?

4- How can the selected SRL components be modified to produce the final AS?

1.4.2 Research Hypothesis

Predicate Argument Structure (PAS) that created from SRL(Semantic Role Labeling) and combined by the use of CST relations can be used to generate good abstractive summary

1.4.3 Research Philosophy

The philosophy behind building AS model based on SRL is to extract the semanticroles to be as a candidate to the abstractive summary AS

1.5 Research Objectives

1.5.1 The research aim

The aim objective is to model a new SRL-CST based technique for accomplishing AS.

1.5.2 The research objectives

- To identify roles in sentence constituents by using SRL (Semantic Role Labeling) technique and then create PAS.

- To develop Case Based Reasoning (CBR) classifier to automatically identify CST (Cross document relation Structure Theory) relation types betweenPAS.
- To investigate the use CST to combine the extracted PAS.
- To generate abstractive summary from combined PAS

1.6 Research Scope

- The research will concentrate on building the SRL-AS model.
- The study focuses only on multi-document abstractive summarization
- A semantic approach for abstractive summarization based on semantic role labeling and CST relations.
- The study focuses on domain specific multi-document summarization where the domain covers news articles related to natural disaster events obtained from the DUC 2002 data set. DUC data set also contains collections of human generated multi-document abstractive summaries, which can be used for evaluation.

1.7 Expected Contribution

The contribution of this study can be identified as follows:

- Automatic extraction of Predicate Argument Structure PASs along with its evaluation.
- Suggested rules to combine PASs using CST.
- SRL-CST based abstractive multi-document summarization.

1.8Thesis Organization

This thesis comprises of 7 chapters which are organised as follows:

Chapter 1 : This chapter overviews the text Summarization in general and the main approaches , Extractive approach and Abstractive one , also it shows that we will going to concentrate on the Abstractive approach , more over the problem statement and research Questions are presented followed by the objectives and the scope of the research as well as the expected contribution.

Chapter 2 : In this chapter, the basic concepts and methods related to our study have been discussed, The chapter starts with the brief introduction to automatic text summarization and provides the past and present works found in the literature. Much discussion was then given for abstractive multi-document summarization task and the methods that are commonly employed for such task. Since this research involves the analysis of multidocument relations, literature works on cross-document relations were also presented. This chapter also reviews all the underlying concepts and techniques that will be used for the proposed methods in this research, such as case-based reasoning, cross-document structure theory.The proposed methods will integrate and combine the advantages of these techniques to achieve the research goals of this study.

Chapter 3 : This chapter shows the methodology that we may use for solving our problem. It contains the generic framework of the research and the steps required to build up the proposed model for automatic abstractive summarization multi-document summarization, it also describes the techniques used to accomplish the research objectives includes semantic role labeling (SRL), extraction of Predicate-Argument Structures (PASs), Cross-document Structure Theory (CST) identification between PASs, combination of PASs based on CST, rank PASs according to position in source text then generate the final abstractive summary.

Chapter 4:This chapter conveys the representation of the dataset from sentence level form to predicate-argument structure form, which is considered as a higher- level of abstraction. This new representation can be processed further in various applications such as text summarization and plagiarism detection. SRL (Semantic Role labeling) is used to identify sentence constituents then, the researcher implements a model to extract the predicate argument structure from the sentences that undergo SRL automatically, the results are compared to a manual predicate-argument structure extraction, a good result has been achieved according to precision and recall values.

Chapter 5:This chapter mentions that the 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 2000) 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..

Chapter 6: In this chapter, we need to combine the predicate argument structure (PASs) according to specific rules suggested by the researcher to get the final summary..

Chapter 7: conclusion , contribution and future work is given in this chapter.

CHAPTERII

Literature Review

2.1 Introduction

With the wide spread use of internet and the emergence of information exploration era, quality text summarization is essential to effectively condense the information. Text summarization is the process of producing shorter presentation of original content which covers non-redundant and salient information extracted from single or multiple documents. Attempts to generate automatic summaries started 50 years ago(Luhn 1958)(Amini et al. 2005)recently, the field of automatic Text Summarization (TS) has experienced an exponential growth due to new technologies.This chapter presents an over view of text summarization in general along with the main classes of text summarization , this chapter also shows the existing significant efforts that have been made in the field of text summarization; and provides the theoretical explanation and fundamental concepts related to it. Moreover, literature reviews on other concepts related to the current study such as semantic role labeling, semantic similarity measures, cross structure document theory and case base reasoning.

2.2 Text Summarization

Automatic text summarization is the summarization through machines, the target of the automatic text summarization is to present a condensed version of text having the key concepts to the user which looks like what human beings do manually.A summary can be defined as a text that is produced from one or more texts, that contains a significant portion of the information in the original text(s), which is no longer than half of the original text(s) (Hovy& Lin 1999). According to(Mani 2001), text summarization is the process of distilling the most important information from a source (or sources) to produce a new version for a particular user (or users)and task (or tasks). Many researchers have defined automatic text summarization from different aspects

One of the researchers defined it as:

" text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than a that" (D.

ways for categorizing text summarization, the purpose and the objective of the summary will Radev et al. 2002).

The other one has defined it as :

"The objective of automatic text summarization is to compress the source document text by extracting its most salient content that satisfies a user's or application needs"(Edmundson 1969).

There are different specify the kind of the final summary produced, figure 2.1 will show the categorize.

.

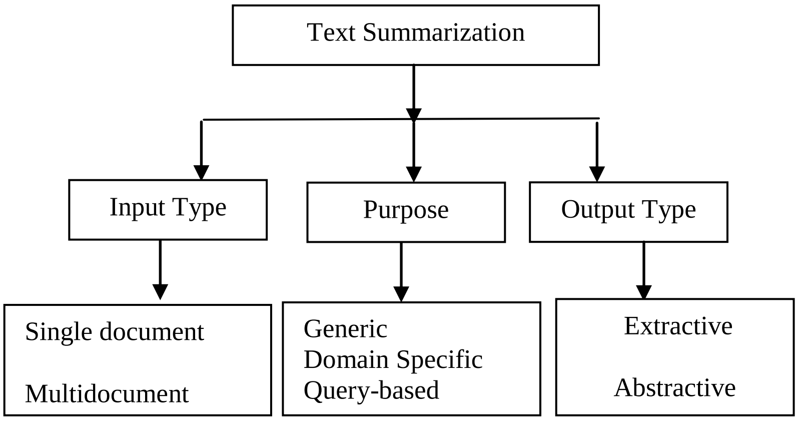


Figure 2.1 : The different categorize of text summarization

2.2.1 Single-document summarization

Single-document summary derives from a single input text (though the summarization process itself may employ information compiled earlier from other texts). Early summarization systems dealt with single document summarization(Hovy & Lin 1996).

2.2.2 Multi-document summarization

A multi-document summary is one text that covers the content of more than one input text, and is usually used only when the Input texts are thematically related(Hovy& Lin 1996).

2.2.3 Generic summarization

The aim of generic summarization is to extract the overall significant information from a document (or a set of documents) regardless of it is topic or domain; i.e. in generic summarization, all documents are viewed as homogenous text and no assumption is made about the domain/topic of source documents. Major work in text summarization revolves around generic summarization(Genest& Lapalme 2011).

2.2.4 Domain specific summarization

Several developments have also been made in various domain-specific summarization systems. For example, summarizing biomedical documents, weather news documents, articles related to terrorist events, finance articles and many more(Radev & McKeown 1998). This type of summarization often requires domain-specific knowledge in the sentence selection process.

2.2.5 Query-based summarization

Query-based summarization extracts important information from document (or documents) that is related to the user's query (or needs). The user queries are usually natural language questions or keywords, related to a particular subject/topic. For instance, the snippet results produced by search engines is an example of a query-based application (Nenkova & McKeown 2012).

2.2.6 Extractive summarization

This type of summarization aims to extract salient sentences from the source documents and concatenated them together to form extractive summary, therefore, the Extraction involves concatenating extracts taken from the corpus into a summary. Most summarization systems that have been developed a deal with extractive summaries(Khan & Salim 2014).

2.2.7 Abstractive summarization

Abstractive summarization consists of understanding the source text by using linguistic and semantic methods to interpret and examine the text. The abstract summary which we are going to focus on is described as an interpretation of an original text. The process of producing involves rewriting (paraphrasing) the original text in a shorter version by replacing wordy concept with shorter ones it is described as an interpretation of an original text. The aim of this type of summarization is to produce a generalized summary which conveys the main information in a concise way. Generally, language generation and compression techniques are required for abstractive summarization(Genest& Lapalme 2011; Khan & Salim 2014).

The study focus on abstractive based domain specific multi-document summarization, in the following sections the researcher goes through a review concerning abstractive multi-document summarization.

Abstractive summarization techniques are broadly classified into two categories: Structured based approach and Semantic based approach. Different methods that use structured based approach are as follows: tree base method, template based method, ontology based method, lead and body phrase method and rule based method. Methods that use semantic based approach are as follows: Multimodal Semantic model, Information item based method, and semantic graph based method.

2.3 Structured Based Approach

Structured based approach encodes most important information from the document(s) through cognitive schemas such as templates, extraction rules and other structures such as tree, ontology, lead and body phrase structure. Different methods used this approach are discussed as follows.

2.3.1. Tree based method

This technique uses a dependency tree to represent the text/contents of a document. Different algorithms are used for content selection for summary e.g. theme intersection algorithm or algorithm that uses local alignment across pair of parsed sentences. The technique uses either a language generator or an algorithm for generation of summary. Related literature using this method is as follows.

The approach proposed in (Barzilay et al. 1999) automatically fuse similar sentences across news articles on the same event. The method uses language generation for producing concise summary. In this approach, first the similar sentences are pre-processed using a shallow parser and then sentences are mapped to predicate-argument structure. Next, the content planner uses theme intersection algorithm to determine common phrases by comparing the predicate-argument structures. Those phrases that convey common information are selected and ordered and some information are also added with it(temporal references, entity descriptions).Finally sentence generation phase uses FUF/SURGE language generator to combine and arrange the selected phrases into new summary sentences. The major strength of this approach is that the use of language generator significantly improved the quality of resultant summaries i.e. reducing repetitions and increasing fluency. However, the approach lacks semantic representation of text as they employed syntactic parsing, and did not report evaluation results.

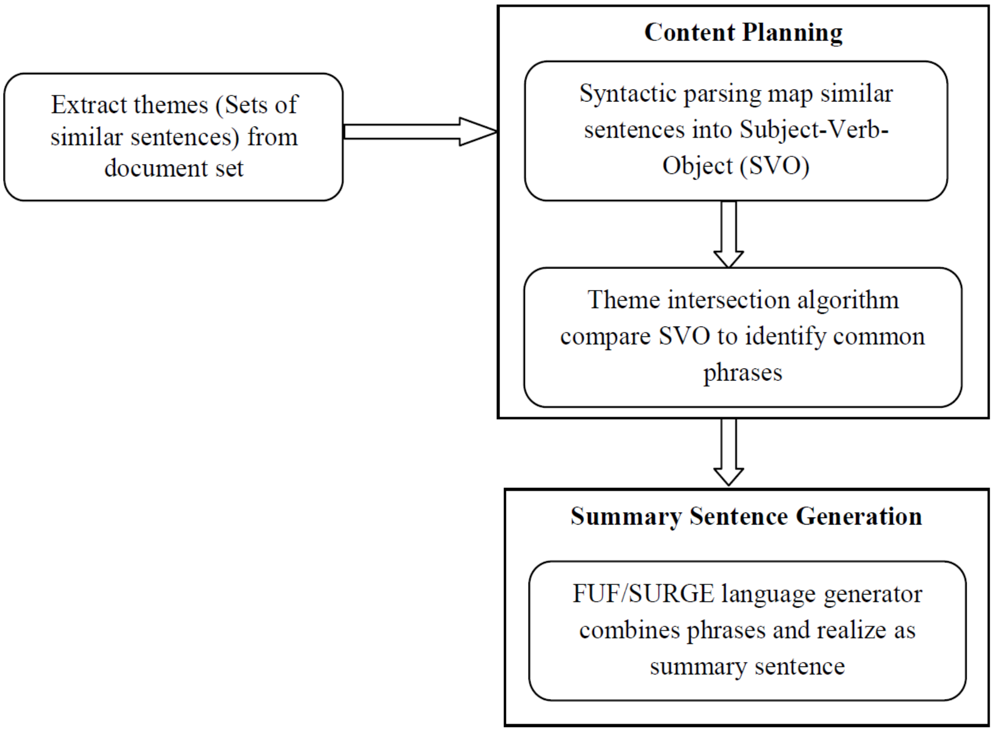

Extract themes (Sets of similar sentences) from document set
Content Planning
Syntactic parsing map similar sentences into Subject-Verb-Object (SVO)
Theme intersection algorithm compare SVO to identify common phrases
Summary Sentence Generation
FUF/SURGE language generator combines phrases and realize as summary sentence

Figure 2.2 :Multi-document summarization based on information fusion (Barzilay*et al.*, 1999)

In other work, sentence fusion (Barzilay & McKeown 2005) integrates information in overlapping sentences to generate a non-overlapping summary sentence. In this approach, first the dependency trees are obtained by analyzing the sentences. A basis tree is set by finding the centroid of the dependency trees. It next augments the basis tree with the sub-trees in other sentences and finally prunes the predefined constituents. The limitation of this approach is that it lacks a complete model which would include an abstract representation for content selection.

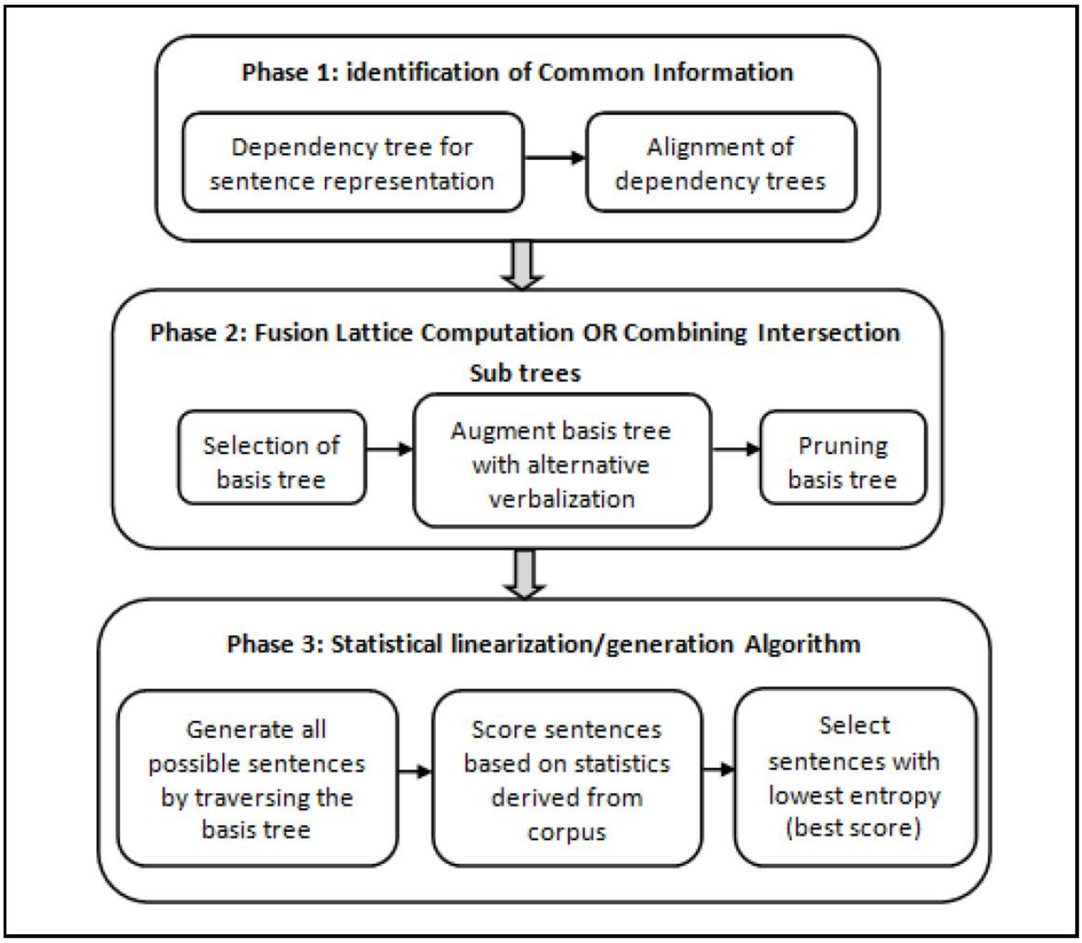

Phase 1: identification of Common Information
Dependency tree for sentence representation
Alignment of dependency trees
Phase 2: Fusion Lattice Computation OR Combining Intersection Sub trees
Selection of basis tree
Augment basis tree with alternative verbalization
Pruning basis tree
Phase 3: Statistical linearization/generation Algorithm
Generate all possible sentences by traversing the basis tree
Score sentences based on statistics derived from corpus
Select sentences with lowest entropy (best score)

Figure 2.3 Phases of sentence fusion (Barzilay and McKeown, 2005)

The goal of first phase in sentence fusion component is to identify the shared information between sentences. The first phase works by representing the input sentences in themes by dependency tree/syntactic parse tree. The dependency trees of pair of sentences are locally aligned based on edge similarity and node similarity; and only high similarity regions (subtrees) are considered as intersection subtrees. The edge is labeled as subject-verb if it connects a subject and verb, and labeled as verb-object if it connects a verb and object node. The node in a dependency tree may be atomic word or phrase (noun phrase). The intersection subtrees (overlapping information of sentence pairs) also called as fragments, are given as input to fusion lattice, the second phase of sentence fusion. The goal of this phase is to combine intersection subtrees. At first step, the basis tree is selected by finding the centroid of the input theme sentences i.e. the sentence which is most similar to the rest of sentences in the input theme or the one that contains most of the fragments. Next, the basis tree is augmented with the information from other input sentences i.e. for each node of the basis tree, the nodes from the other input trees that are aligned with a given node are determined and their corresponding verbalizations (word/phrase) are recorded. In last step, basis tree is pruned off by removing sub-trees which are not part of intersection. The last phase of sentence fusion aims to linearize the fusion lattice or to generate sentence from basis tree. At first step, basis tree or fusion lattice is traversed to generate all possible sentences. Next, the likelihood of sentences are scored based on statistics derived from corpus and finally the sentence with best score (lowest entropy) is chosen as verbalization of basis tree. The limitation of this approach is that it lacks a complete model, which would include semantic representation of text for content selection.

2.3.2 Lead and body phrase method

This focuses on the phrases that has got same syntactic head chunk in lead and body sentences. Here the same chunks are searched in lead and body sentences (Tanaka et al. 2009). Then these phrases are aligned using similarity metric. If the body phrase has rich information and has same corresponding phrase then substitution occurs. But if body phrase has no counterpart then insertion takes place as shown in figure 2.4. The potential benefit of this method is that it found semantically appropriate revisions for revising a lead sentence but it has a drawback of rewriting the sentences.

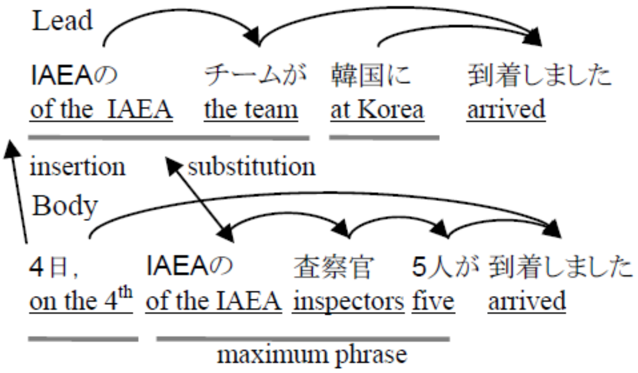


Figure 2.4 Approach of revising lead sentence (Tanaka, Kinoshita et al. 2009)

2.3.3 Information item based method

In this method, the contents of summary are generated from abstract representation of source documents, rather than from sentences of source documents. The abstract representation is Information Item, which is the smallest element of coherent information in a text.

A framework for multi-document abstractive summarization of news is presented by (Genest & Lapalme 2011), as shown in figure 2.5. The framework consists of following modules: Information Item retrieval, sentence generation, sentence selection and summary generation. In Information Item (INIT) retrieval, first syntactic analysis of text is done with parser and the verb’s subject and object are extracted as shown in figure 2.6.

In sentence generation module, a sentence is directly generated from INIT using parse tree of the sentence from which INIT is taken and the language generator, NLG realizer SimpleNLG(Gatt & Reiter 2009). Sentence selection module ranks the generated sentences based on their average Document Frequency (DF) score. Finally, a summary generation step account for the planning stage and include dates and locations for the highly ranked generated sentences.

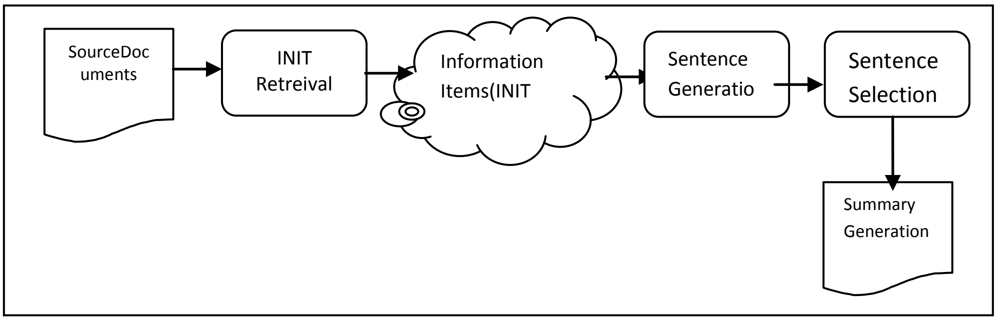


Figure 2.5 Information item based method for abstractive summarization (Genest and Lapalme 2011)

Original Sentence At least 25 bears died in the greater Yellowstone area last year, including eight breedingagefemales killed by people.

Information Items

1. bear – die – null (greater Yellowstone area, last year)
2. person – kill – female (greater Yellowstone area, last year)

Generated Sentences

1. 25 bears died.
2. Some people killed eight breeding-age females.

Selected Generated Sentence as it appears in the summary

1. Last year, 25 bears died in greater Yellowstone area.

Figure 2.6 Example of Information item based method

However, this study did not compare INITs with similar meanings nor ranking was considered based on INITs. Moreover, INITs belong to single sentence were not grouped to acquire the full semantic of the sentence and the linguistic quality of generated summary was low due to incorrect parse.

2.4 Semantic Based Approach

Semantic based approach aims to produce abstractive summary from semantic representation of document text. Semantic representation of text include predicate argument structure representation extracted through semantic role parsing (arguments of the predicate are augmented with semantic roles), ontological representation of document, template representation of topic extracted through information extraction (IE) systems.

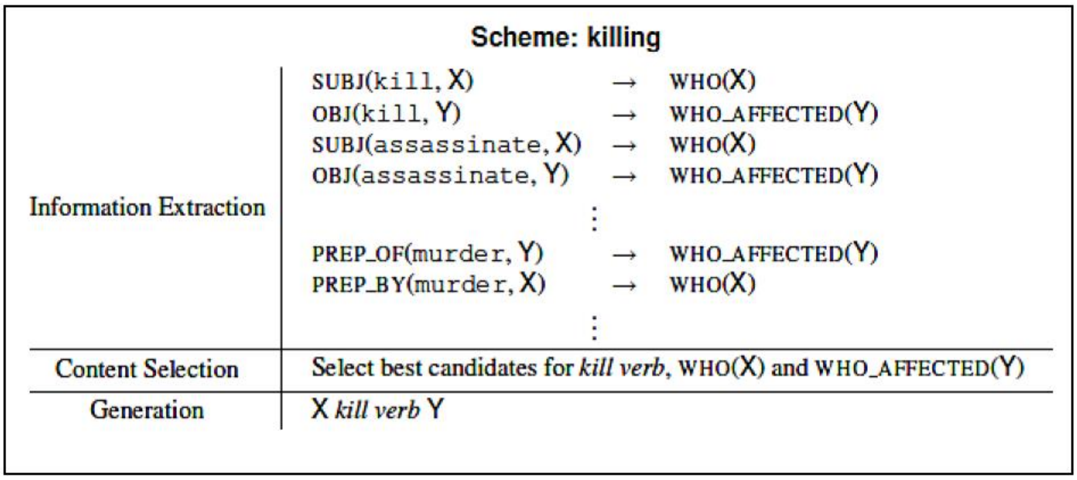
The semantic based approaches discussed in the literature are template based method, ontology based method and graph based method. These methods are discussed as follows:

2.4.1 Template based method

In this method, the source document(s) is represented by a template. Numerous linguistic patterns or extraction rules are designed to match with the source text, and discover the text snippets that will be mapped into template slots.

These text snippets are indicators of the summary content. In this method, the documents to be summarized can also be represented in terms of categories and a list of aspects. Content selection module then selects the best candidate among the ones generated by information extraction rules to answer one or more aspects of a category. Finally, generation patterns are used for generation of summary sentences. One of Related literature employing this method is discussed as follows:

Template based methodology demonstrated by(Genest & Lapalme 2012) generates short and well written abstractive summaries from clusters of news articles on same event. The methodology is based on an abstraction scheme as shown in figure 2.7. The abstraction scheme uses a rule based information extraction module, content selection heuristics and one or more patterns for sentence generation. Each abstraction scheme deals with one theme or subcategory. In order to generate extraction rules for abstraction scheme, several verbs and nouns having similar meaning are determined and syntactic position of roles is also identified. The information extraction (IE) module finds several candidate rules for each aspect (What, When, Where, Why, Damages) of the category (Accidents and Natural Disasters, Attacks).



Scheme: killing

Information Extraction	SUBJ(kill, X)	→	WHO(X)
	OBJ(kill, Y)	→	WHO_AFFECTED(Y)
	SUBJ(assassinate, X)	→	WHO(X)
	OBJ(assassinate, Y)	→	WHO_AFFECTED(Y)
		⋮	
	PREP_OF(murder, Y)	→	WHO_AFFECTED(Y)
	PREP_BY(murder, X)	→	WHO(X)
		⋮	
Content Selection	Select best candidates for *kill verb*, WHO(X) and WHO_AFFECTED(Y)		
Generation	X *kill verb* Y		

Figure 2.7 Rule based abstraction scheme for killing (Genest and Lapalme 2012)

Based on the output of the IE module, the content selection module selects the best candidate rule for each aspect and passed it to summary generation module. This module exploits generation patterns designed for each abstraction scheme, in order to decide the structure of the generated sentence. Next, sentence structure and words are given as input to SimpleNLG realizer, which give the generated sentence. The main drawback of this methodology is that all the rules and patterns are written by hand, which is tedious and time consuming.

2.4.2 Ontology based method

Many researchers have made effort to use ontology (knowledge base) to improve the process of summarization. Most documents on the web are domain related because they discuss the same topic or event. Each domain has its own knowledge structure and that can be better represented by ontology. One of related literature using this method is discussed as follows:

The fuzzy ontology with fuzzy concepts is introduced for Chinese news summarization (Lee et al. 2005) to model uncertain information and hence can better describe the domain knowledge. In this approach, first the domain expert pre-defines the domain ontology (for news events) and the Chinese news dictionary, as shown in figure 2.8. The retrieval agent retrieves weather news from internet and store in news corpus for processing. Next, during the document pre-processing phase, the significant terms

(Nouns and Verbs) are extracted based on POS tags defined in Chinese news dictionary and the term frequencies are derived from news corpus.

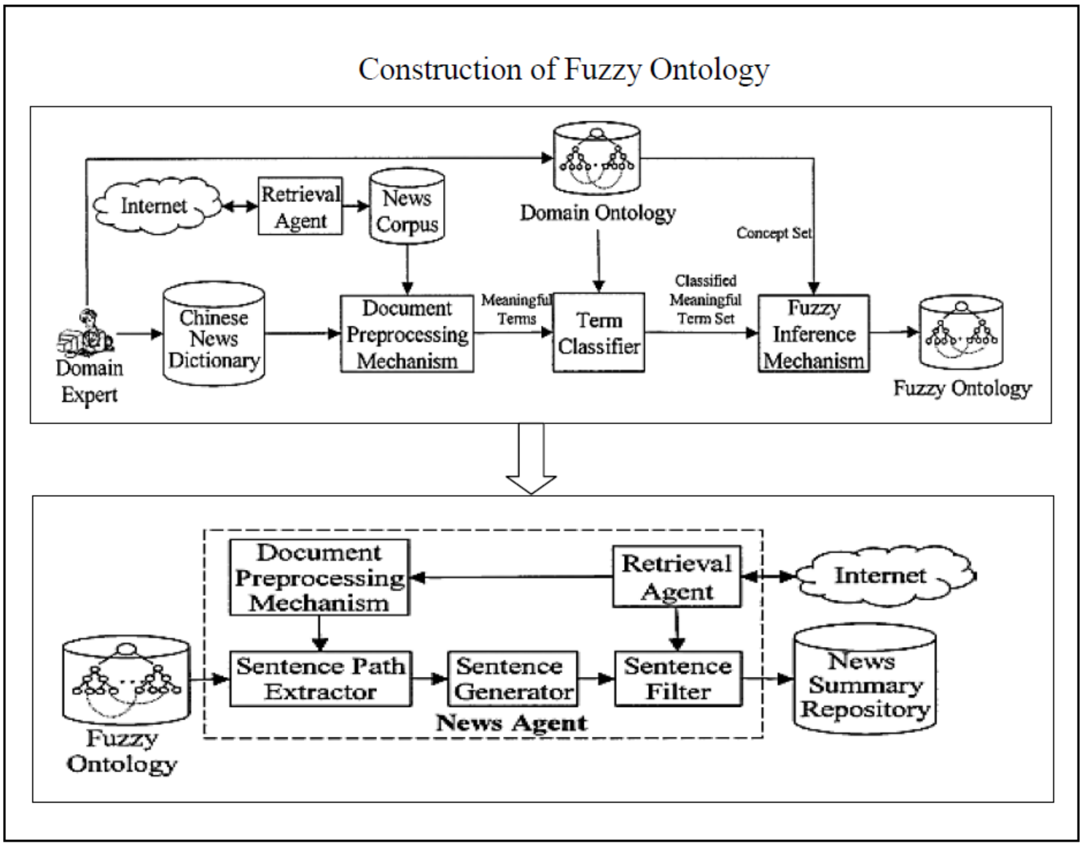


Figure 2.8 Fuzzy ontology equipped with News Agent for news summarization(Lee et al. 2005) .

Once the significant terms are extracted, the term classifier classifies the meaningful terms on the basis of events of news. The fuzzy inference phase produces a fuzzy ontology by associating membership degrees to each concept in domain ontology. A set of membership degrees belong to each fuzzy concept is linked with different domain ontology events. News summarization is done by news agent equipped with fuzzy ontology. The document pre-processing and retrieval agent components in news agent perform the same tasks as discussed earlier.

The news agent comprises of three main components: (a) sentence path extractor, (b) sentence generator and (c) sentence filter. The sentence path extractor takes meaningful terms and all fuzzy concepts as input and extracts all possible sentence paths from fuzzy ontology. The sentence generator will generate set of sentences by exploiting class layer of fuzzy ontology, and finally the sentence filter component of news agent will remove noisy sentences and generate set of summarized sentences. The benefit of this approach is that it exploits fuzzy ontology to handle uncertain data that simple domain ontology cannot. However, this approach has several limitations. First, domain ontology and Chinese dictionary has to be defined by a domain expert which is time consuming. Secondly, this approach is limited to Chinese news, and might not be applicable to English news.

2.4.3 Graph-based method

In this method, a document to be summarized is represented rich Semantic Graph (RSG), exploits some heuristics to reduce the semantic graph, and then the reduced semantic graph is used in abstractive summary generation. The abstractive approach proposed by (Moawad & Aref 2012) operates in three phases as shown in figure 2.9.

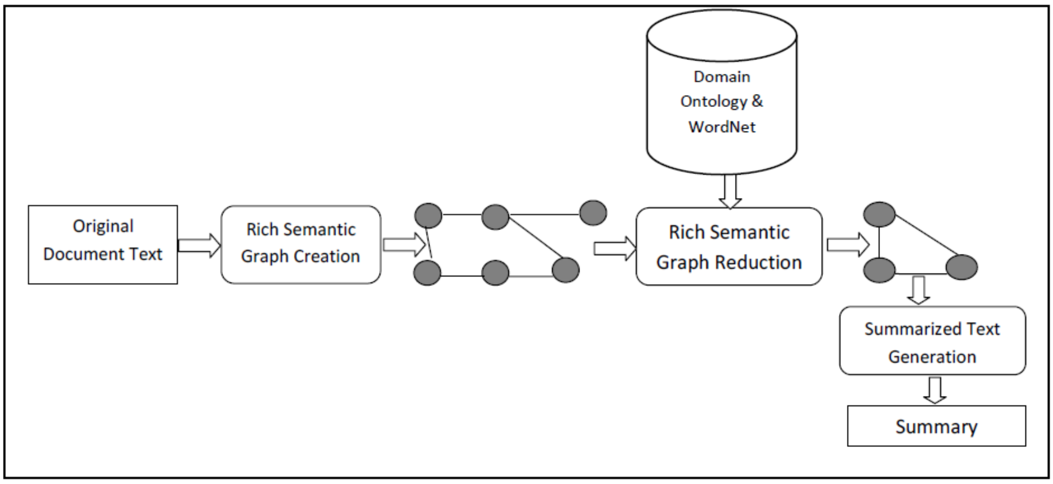


Figure 2.9 Abstractive summarization with semantic graph(Moawad and Aref 2012)

.

The first phase represents the input document to be summarized by Rich Semantic Graph (RSG). In RSG, the graph nodes represent nouns and verbs in the input document and the edges correspond to semantic and grammatical relations between them. RSG is an ontology based representation i.e. the graph nodes are the instances of noun and verb classes in domain ontology.

In the second phase, a set heuristic rules exploiting hypernym and holonym semantic relations from WordNet, are applied to RSG and mitigate it by deleting, replacing or merging the graph nodes, resulting in a reduced rich semantic graph (RRSG).

Finally, the summary generation module as shown in Figure 2.10, generates the summary in four steps: (a) text planning, (b) sentence planning, (c) surface realization and (d) evaluation. The goal of text planning step is to choose the relevant/suitable content that will be incorporated in the generated text.

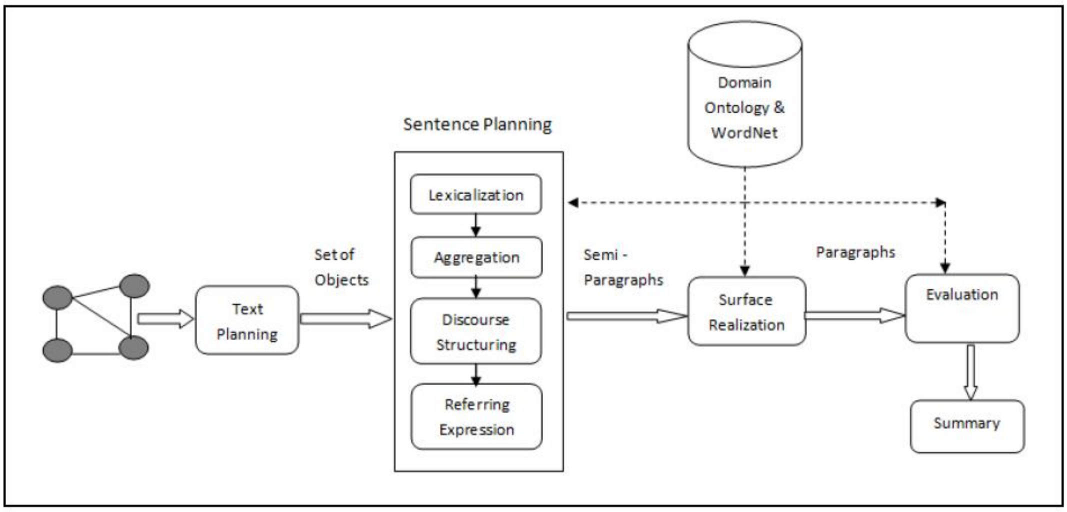


Figure 2.10 Process of summary generation module (Moawad and Aref 2012)

In order to perform the desired task in text planning, all the graph objects i.e. noun and verb objects are chosen and given to the sentence planning step, which produces semi-paragraphs. The aim of sentence planning step is to make the text fluent and understandable, and is achieved by employing four core processes: lexicalization, aggregation, discourse structuring, and referring expressions. Lexicalization process aims to access the WordNet to select the appropriate synonyms for each verb and noun object in order to generate the target text.

Next, process of discourse structuring builds pseudo-sentences (initial form of sentences) that contain the selected object (noun and verbs) synonyms. Once the pseudo-sentences are generated, the aggregation process combines them into semi-paragraphs by employing two processes: predicate grouping and subject grouping. Predicate grouping aims to combine clauses with same predicates, while subject grouping aims to combine clauses with same subject. Discourse relations from domain ontology are also exploited to connect pseudo-sentences into semi-paragraphs.

Next, in referent expression process, each pseudo-sentence in the semi-paragraph is scanned to identify the repeated subject and store it in a list. Then, the repeated subject in replaced with suitable pronoun in every pseudo-sentence of a semi-paragraph except the subject of first pseudo-sentence.

After the semi-paragraphs are generated, the surface realization step utilizes SimpleNLG to transform them into grammatically correct paragraphs (tenses of word, adding punctuation like semi-colon). Finally, in the evaluation step, the paragraphs are ranked based on two features: coherence relations between sentences of paragraph and the most frequent word synonyms in paragraph.

However, this approach has some drawbacks. First, it relies on domain expert for constructing domain specific ontology, which is limited to a particular domain; and in case the domain changes, the ontology will need to be rebuilt and therefore the semantic graph will be re-constructed. Secondly, this approach is applied only to single document and did not report any evaluations.

2.5 Related Techniques used in proposed Methods

In this study the researcher uses different techniques to accomplish the work these techniques are as follows :

2.5.1 Semantic Role Labeling (SRL)

SRL is a task in natural language processing (NLP) consisting of detection of the semantic arguments associated with the predicate or verb of a sentence and their classification to their specific roles , more over it is the underlying relationship that a participant has with the main verb in the clause (Genest & Lapalme 2012), also known as semantic case, thematic role, theta role (generative grammar), and deep case (case grammar). The goal of SRL is to discover the predicate argument structure of each predicate in a given input sentence(Suanmali et al. 2011) . According to(Khan et al. 2015a) the task of SRL is to find all arguments for a given predicate in a sentence and label them with semantic roles.

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 . As the definition of the PropBank and CoNLL- 2004 shared task(Collobert, Weston et al. 2011)there are six different types of arguments labeled as A0-A5 and AA. These labels have different semantics for each verb as specified in the PropBank Frame files. In addition, there are also 13 types of adjuncts labeled as AM-adj where adj specifies the adjunct type as shown in table 2.1 . 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 (Johansson & Persson 2009). 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 (Màrquez et al. 2008).

.SRL aims to identify the constituents of a sentence, together with their roles with respect to the sentence predicates. In this study Predicate Argument Structure(PAS) extracted from sentences is used as semantic representation for sentences in documents collection therefore it is used as a representation for our dataset, we use DUC 2002 dataset , for the SRL we use SENNA toolkit(Aksoy et al. 2009) .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).

Table 2.1 : Representation of Core Arguments and Adjunctive Arguments(Kumar et al. 2013)

Core Arguments		**Adjunctive Arguments**
	verb	*ArgM-ADV*adverbial modification
0	subject	*ArgM-DIR* direction
1	object	*ArgM-DIS* discourse marker
2	Indirect object	*ArgM-EXT*extent marker
3	Start point	*ArgM-LOC*location
4	End point	*ArgM-MNR*manner
5	Direction	*ArgM-MOD*general modification
		*ArgM-NEG*negation
		*ArgM-PRD*secondary predicate
		*ArgM-PRP*purpose
		*ArgM-REC*reciprocal
		*ArgM-TMP*temporal marker

The process of SRL operates in three steps: In the first step, the sentence is parsed and represented by a syntactic parse tree. The nodes of the parse tree represent syntactic categories such as NP, VP and PP where NP stands for noun phrase, VP for verb phrase and PP for preposition phrase. The leaves of the parse tree represent the tokens (words) of the sentence, while their corresponding part-of-speech tags (e.g. NN (Noun, singular or mass), VBZ or VBP (present tense verb), etc.) appear one level above the leaf nodes

2.5.2 Cross document Structure Theory (CST)

The study on multi-document relations was pioneered by Radev (D. R. Radev et al. 2002). Radev introduced the CST model (Cross-document Structure Theory). The general schema of CST is shown in figure 2.11. Its fundamental idea is that documents which are related to the same topic usually contain semantically related textual units. These textual units can be words, phrases, sentences, or the documents itself. In our work, we investigate only the semantic relations between sentences.

Up To now, cross document relations and discourse relations have benefit various NLP applications such as text summarization(D. R. Radev et al. 2002; Adilah & Zahri 2012; Kumar et al. 2012); (Kumar et al. 2013) .

In text summarization, discourse relations are used to produce a best ordering of sentences ina document, and remove redundancy from generated summaries.

The famous well known works isCST based text summarization(Zhang et al. 2002)).In this work, sentences with most relations inthe documents are considered to be important. They proposed an enhancement of textsummarization by replacing low-salience sentences with sentences having maximum numbers ofCST relations.

Table 2.2: Relation Types and their meaning(Zhang et al. 2002)

	Relationship	Description	Textspan 1(S1)	Textspan2(S2)
1	identity	location	termtoday.	termtoday.
2	Equivalence (Paraphrase)	otextspanshavethesameinformation content	DerekBellisexperiencinga resurgence in hiscareer.	Derek Bell is having a ”comeback year.”
3	Translation	ameinformationcontentindifferent languages	Shoutsof“Vivala revolucion!”echoed throughthenight.	The rebels could be heard shouting, “Longlivetherevolution”.
4	Subsumption	1containsallinformationinS2,plus additionalinformationnotin S2	With3winsthisyear,GreenBayhas thebestrecordin theNFL.	GreenBayhas 3winsthisyear.
5	Contradiction	Conflictinginformation	Therewere122peopleonthedowned	126peoplewereaboardtheplane.
6	Historical Background	giveshistoricalcontexttoinformationi nS2	Thiswasthefourthtimeamemberof theRoyalFamilyhasgottendivorced.	The Duke of Windsor was divorced fromtheDuchessofWindsoryesterday.
7	Citation	S1 explicitlycitesdocumentS2	AnearlierarticlequotedPrinceAlbert assaying“I never gamble.”	PrinceAlbertthenwentontosay,“I nevergamble.”
8	Modality	1presentsaqualifiedversionofthein- rmationin S2,e.g.,using“allegedly”	Sean“Puffy”Combsisreportedto own severalmultimilliondollarestates.	Puffy owns four multimillion dollar homesin theNew Yorkarea.

Table 2.2: Relation Types and their meaning Cont (Zhang et al. 2002)

D	Relation ship	**Description**	**Textspan 1(S1)**	**Textspan2(S2)**
9	Attribution	S1 presentsanattributed versionofinformation inS2, e.g.using“According toCNN,”	AccordingtoatopBushadvisor, the President wasalarmed atthenews.	ThePresidentwasalarmed tohearof hisdaughter’slow grades.
10	Summary	S1summarizesS2.	TheMetswontheTitleinseven games.	After a gruelingfirstsix games, the Metscamefrombehind tonighttotake theTitle.
11	Follow-up	S1 presents additional information whichhashappenedsinceS2	102casualtieshavebeen reportedintheearthquakeregion.	ofar, nocasualtiesfromthequake have beenconfirmed.
12	ndirectspeech	S1indirectlyquotessomethingwhich wasdirectlyquoted in S2		“I’llpersonally guaranteefreeChalupas,”Mr.Cuban announced tothe crowd.
13	Elaboration Refinement)	S1 elaborates or provides details of someinformationgivenmore generally inS2	0%ofstudentsareunder25;20%arebetween26and30;therestareover30.	MoststudentsattheUniversityare under30.
14	Fulfillment	S1assertstheoccurrenceofanevent predictedin S2	Aftertravelingto AustriaThursday,Mr. Green returnedhometo New York.	Mr.Greenwillgo toAustriaThursday.
13	Elaboration Refinement)	S1 elaborates or provides details of someinformationgivenmore generally inS2	0%ofstudentsareunder25;20%arebetween26and30;therestareover30.	MoststudentsattheUniversityare under30.

Table 2.2: Relation Types and their meaning cont.(Zhang et al. 2002).

ID	Relation ship	**Description**	**Textspan 1(S1)**	**Textspan2(S2)**
16	ReaderProfile	S1andS2providesimilarinformation writtenfora differentaudience.	TheDurian,afruit usedinAsian ,hasa strongsmell.	Thedishis usuallymadewithDurian.
17	hangeofperspective	Thesameentitypresents a differing opinion orpresents afactinadifferent light.	Giulianicriticizedthe Officer’sUnion as too demanding”incontracttalks.	Giuliani praised the Officer’sUnion, which provideslegalaidandadvice to members.
18	Overlap (partial equivalence)	S1providesfactsXandYwhile S2providesfactsXandZ;X,Y ,andZshould allbenon-trivial.	Theplanecrashedintothe 25thfloorof thePirelli buildingin downtownMilan.	Asmalltouristplanecrashedintothe tallestbuildinginMilan.

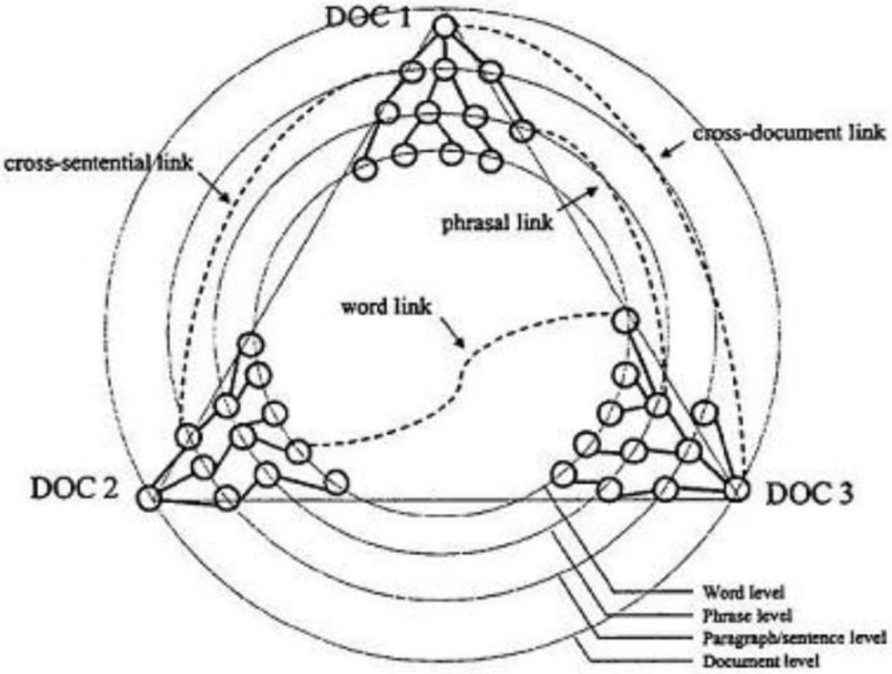


Figure 2.11 : CST general schema(Radev 2000)

Table 2.3: CST Relations used in this work

CST Relation	Description
Identity	The same text appears in more than one location
Subsumption	1 contains all information in S2 , plus additional information not in S2
Description	S1 describes an entity mentioned in S2
Overlap	S1 provides facts X and Y while S2 provides facts X and Z ; X,Y, and Z should all be non-trivial.

The ability to automatically identify the CST relations from un-annotated text could be useful for applications related to multi-document analysis. For instance, a number of works have addressed the benefits of CST for summarization task. However these works relies on text documents which were already annotated with CST relations. Thus the need for automation is deemed necessary.

Majority of the CST-based works tracked the effects of individual CSTrelationships to the summary generation The famous well known works isCST based text summarization(Zhang et al. 2002). In this work, sentences with most relations inthe documents are considered to be important. They proposed an enhancement of textsummarization by replacing low-salience sentences with sentences having maximum numbers ofCST relations. Many other researchers investigate CST as summarized in table (2.3) .

Table 2.4 : Examples of Approaches used for identification of CST relations

Authors	Approaches used for identification of CST	Types of CST relations used	eatures used for identification of CST relation	Machine learning technique have been used	Disadvantages in this approach
(1) Y.J. Kumar et al. pplied Soft Computing 21 (2014) 265–279	identify the relations etween sentences directly from un-annotated ocuments by benefits from eviously identified similar ase and this is done using CBR	Four types CST relations,namely Identity, Subsumption, Description nd Overlap; as they cover ost of the CST relations in the CST model	Uses 5 features i.e Cs(cosine similarity),WO(word overlap),Lt(length type),NP(noun phrase),VP(verb phrase)	He propose a supervised learning method based on ase based reasoning (CBR) technique which is optimized using genetic learning algorithm(CBR-Gent)	The Identification of CST Relation between sentences applied only on 4 CST Relation out of 18 Relation
Yogan Jaya Kumar laomieSalim, BasitRaza ross-document structural lationship identification sing supervised machine Learning	Uses CSTBank dataset which already annotated nd the use SVM to classify NewSentences to their suitable CST relations	entity,subsumptionDescription,overlap,	CS,WO,LT,NP,VP	SVM,NN,CBR	lo scaling for the features The features elevance is regarded equally for the 5 features
(2)Z. Zhang, S. Blair-oldensohn, D.R. Radev, owards CST-enhanced ımmariza- tion, in: Proc. AAAI/IAAI, 2002, pp. 439–446.	he identification of CST is one manually by subjects	Relationship Elaboration / Refinement Equivalence Description Historical Background Follow-up Subsumption Contradiction Attribution Identity Indirect speech Fulfillment Modality Summary eader Profile 1 Change of Perspective 1 Translation	e mention the overall score based of features but not identified them.	They use MEAD summarizer with nput clusters of annonated sentences with the CST Relationship connectivity	he major limitation of this work is that he CST relations need to be manually nnotated by human experts; which is a drawback for an automatic summarization system

Table 2.4 : Examples of Approaches used for identification of CST relations Cont.

Authors	**pproaches used for lentification of CST**	**ypes of CST relations used**	**Features used for identification of CST relation**	**Machine learning technique have been used**	**Disadvantages in this approach**
Identifying Multidocument Relations Erick GalaniMaziero, Maria Lucía del Rosario Castro Jorge, ThiagoAlexandreSalgueiroPardo NúcleoInterinstitucional de LingüísticaComputacional (NILC) Instituto de CiênciasMatemáticas e de Computação, Universidade de São Paulo Av. Trabalhador São-carlense, 400. P.O.Box. 668. 13560-970 - São Carlos/SP, Brazil {erickgm,mluciacj,taspardo}@icmc.usp.br	'hey used CSTTool for xt anntation with CST Relations	Identity Modality Equivalence Attribution Translation Summary Subsumption Follow-up Contradiction Elaboration Iistorical -background Indirect speech Citation Overlap	difference in length of sentences (in number of words), percentages of common words in the sentences, position of each sentence in the text that it belongs to, a flag indicating whether a sentence is shorter than the other, a flag indicating whether the sentences identical, and the number nouns, proper nouns, adverbs, adjectives, verbs and numerals in each sentence	They use J48 for decision tree generation, which belongs to the mbolic paradigm. -- Naïve-bayes -stratified ten-fold cross-validation technique for training and testing	CST Relation need better identification

In the next section, the fundamental concept of our proposedapproach will reviewed (i.e. the case base reasoning approach), which will be used in this study toclassify the cross-document relations in texts.

2.5.3 Case Base Reasoning (CBR)

Case Based Reasoning (CBR) is a family of artificial intelligence techniques, based on human problem solving paradigm(Kumar et al. 2012). 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.

When a new case (problem) is received, the CBR model will first retrieve themost similar cases from the case base (where previous solved cases are stored) andthe solution from the retrieved cases will be reused for the new case. If no similarcases are found in the case base, the solution for the new case will be revised andretained into the case base as a newly solved case.

CBR usually requires much less knowledge acquisition and does not requirethe extraction of domain model as it relies on the collection of past experiences (Yao & Li 2010). Another interesting characteristic of this method is that it is capable toadapt new cases to its case base, whereby this method does not require retraining ofdata which is necessary for most supervised machine learning techniques (Malhotra 2011).

2.5.3.1 CBR lifecycle

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)(Aamodt& Plaza 1994).

The four processes are illustrated in Figure 2.12

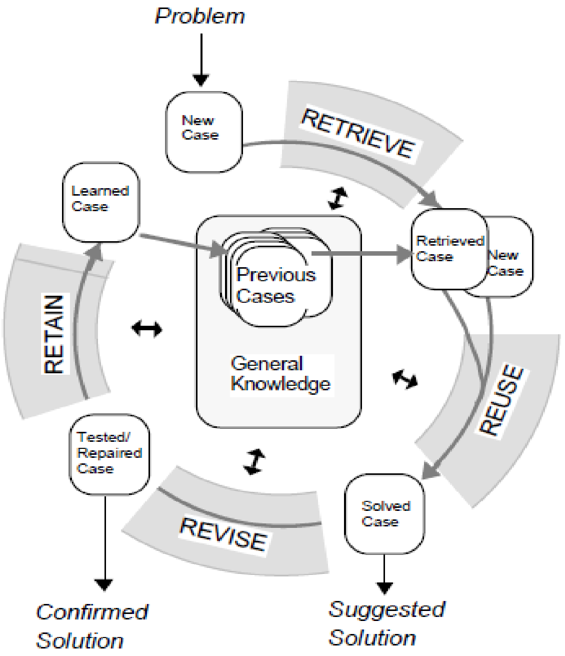


Figure 2.12: CBR life cycle(Aamodt and Plaza 1994)

2.6 Evaluations Measures

In order to evaluate the system generated summary, standard evaluation measures need to be applied. The two standard evaluation metrics, Recall-Oriented-Understudy for Gisting Evaluation (ROUGE) (Lin 2004)and Pyramid (Nenkova & Passonneau 2004)have been widely used in the context of evaluation of text summary.

ROUGE-N is one of the variants of ROUGE measure that determines an n-gram recall between a system summary and a set of reference summaries. On other hand, pyramid score for a new peer summary (system summary) is the ratio of sum of weights of its peer summary content units (SCUs) to the average SCUs in the model summary.

2.6 Chapter Summary

In this chapter the basic concepts and methods related to our study have been discussed , The chapter starts with the brief introduction to automatic text summarization and provides the past and present works found in the literature. A lot ofdiscussion was then given for abstractive multi document summarization task and the methods that are commonly employed for such task. Since this research involves the analysis on multi document relations, literature works on cross-document relations were also presented . This chapter also reviews all the underlying conceptsand techniques that will be used for the proposed methods in this research, such ascase-based reasoning, cross document structure theory .The proposed methods will integrate and combine the advantagesof these techniques to achieve the research goals of this study. The next chapter(Chapter 3) will present the methodology that will be followed to meet the research goals.

CHAPTERIII

Research Methodology

3.1 Introduction

This chapter presents the methodology used in this research. It contains the genericframework of the research and the steps required to build up the proposed model for automatic abstractive summarization multi-document summarization, it also describesthe techniques used to accomplish the research objectives. This study includes semantic role labeling (SRL), extraction of Predicate Argument Structures (PASs), Cross document Structure Theory (CST) identification between PASs, combination of PASs based on CST, rank PASs according to position in source text then generate the final abstractive summary . This chapter proceeds as follows: Section 3.2explains the research design of this study. Section 3.3 presents the researchoperational framework and finally the chapter summary in Section 3.4.

3.2 Research Design

The research design gives detailed steps of the research.Since the research is a process; we must organize the steps of this process to meet the research goals. Each process must be well understood so that we can move to the next stage.

As we convey in chapter1, we have four goals(objectives) to be undertaken with the aim of finding answers to our research questions, these includes identification of the roles in sentence constituents by using SRL(Semantic Role Labeling) techniques and then create PAS, develop case base reasoning (CBR) classifier to automatically identify CST(Cross document relation Structure Theory) relation types between PAS, to investigate the use of CST to combine the extracted PAS and to generate the abstractive summary from the combined PAS.

The research design of this study combined several techniques in order to establish a new model of semantic abstractive multi-document summarization based on CST.

These techniques consist of semantic role labeling,extraction of PAS,and cross document relation identification between PASs.This study consists of five phases with each phase containing a number ofsteps, the phases are briefly mentioned as follows:

The first phase is a preliminary study that includes literature review which covers previous and recent works related to this study. Problem formulation and identification of existing techniques are also covered in this phase. A part of this phase is dedicated to collect the data required for this study; DUC data 2002 will be used as an evaluation data. This phase will include data pre-processing such as breaking the document into sentences, tokenization, stop word removal and word stemming.

The second phase introduces semantic approach for multi-document abstractive summarization using semantic role labeling (SRL) technique. SRL identifies semantic representation (predicate argument structures) from document text automatically for generating a good abstractive summary.The proposed approach extracts predicate argument structures (PASs) from the document text, and build semantic similarity matrix from the pair wise similarities of predicate argument structures (PASs), computed using Jiang's semantic similarity measure (Jiang & Conrath 1997).

The third phase investigates the identificationof cross document relation (CST) between PASs, therefore we develop a case base reasoning (CBR) classifier to automatically identify the existence of cross document relations (CST) betweeneach pair of PASs.

The fourth phase concerns with the score procedure which given to each PAS as an evaluation scheme , therefore according to number of relation types that each PAS holds a score is given to each PAS , the scored PASs were then ordered using document NO and the sentence position NO in that document , lastly the selected higher scored PASswere ordered to constitute the final summary.

The fifth phase of this study is writing up the thesis that combines the problem statement, objectives, scopes of this study and as well as the literature reviews related to the abstractive text summarization. The thesis also described the methodology employed in the research. It also contains the experiments, results and the analysis of the conducted research. Finally, the conclusion and suggestions for further works are also included in this report. .

3.3 Research Frame Work

The frame work provides a well-defined, systematic and structured guidance towards conducting the research at each stage of the research process, figure 3.1 illustrates the research operational frame work.

Our Research Frame work consist of five phases phase 1: Preliminary Study and Data Preparation which related to Problem formulation, phase 2:Extraction of predicate argument structure after semantic role labeling, phase 3: Cross document relation identification using CBR approach, phase 4: Semantic approach of multidocument abstractive summarization using SRL and CST, phase 5:writing up the thesis chapters.

In the following sections we go through each phase in detail.

3.3.1 Phase 1: Preliminary Study and Data Preparation

Preliminary Study This phase composed of three main elements: The planning and literature review, collecting DUC data for evaluation and data pre-processing. The planning and literature review consisted of three steps: formulation of problem, review of literature; and study and identification of existing techniques. Problem formulation step determine such issues in multi-document abstractive text summarization that does not have solutions or the available solutions still have chances for improvement. The literature review studies the research work related to the current multi-document abstractive text summarization methods in order to analyze the summarization techniques used in these methods and discussed their strengths and shortcomings. Throughout the literature review, relevant information to our research work will be studied.

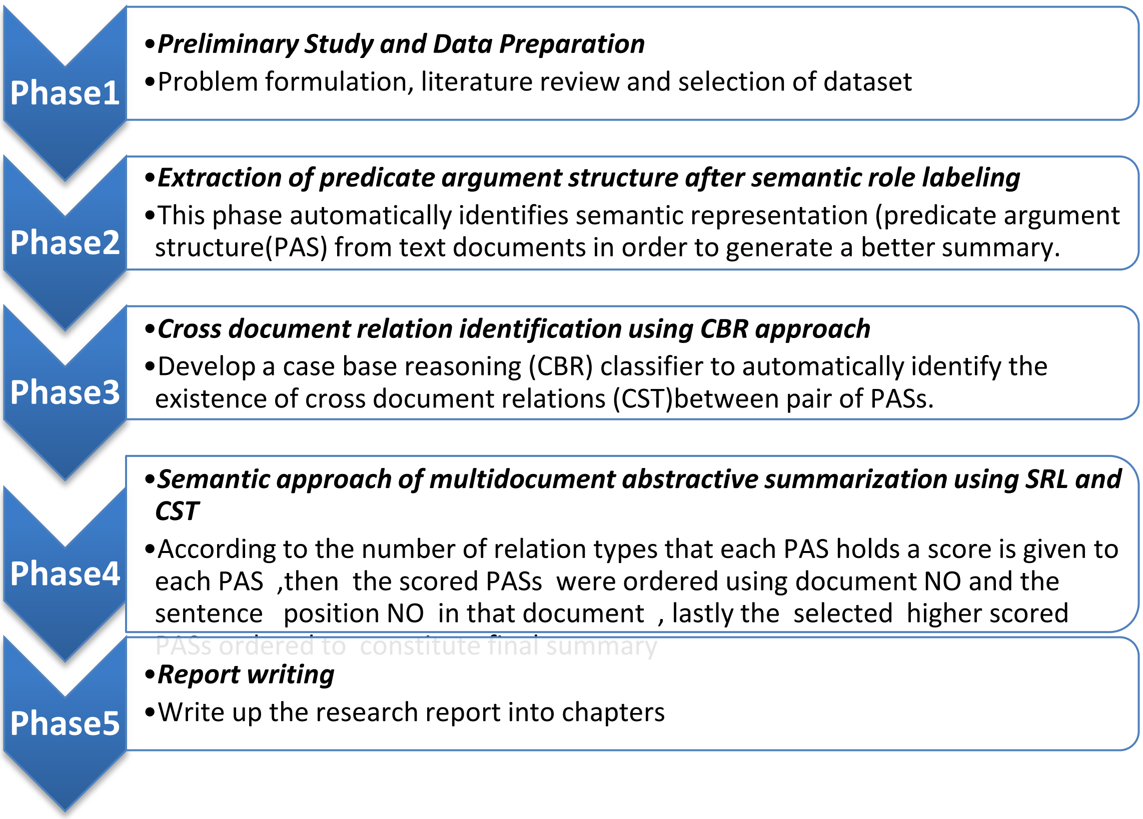
Data Gathering:Another important part in this initial phase is data gathering. Data gathering involves selection of data sets to be used for the purpose of research evaluation. There are two main data sets which we will require for this study; one is for evaluating our proposed methods for cross-document relation identification and the other is for the evaluation our proposed summarization models. For the crossdocumentrelation identification, we exploit the publically available CSTBank corpus (Radev et al. 2004)– a corpus consisting clusters of English news articles annotatedwith cross-document relations. Using the datasets from CSTBank, we will be able toprepare our training and testing data which comprises of sentence pairs with its classlabel (cross-document relation).

Next, to test our summarization methods, we use the DocumentUnderstanding Conference (DUC) data collection DUC 2002 (Over 2002). DUC stands for Document Understating Conference and is a benchmark data set widely employed in the field of text summarization research. It consists of documents along with their corresponding human made summaries. The DUC data is collected from famous newswires used by most researchers in automatic text summarization. The evaluation data of 59 document sets (multi-documents) were used in DUC 2002. Each document set contains documents, abstracts of single document, and extracts/abstracts of multi-document with sets explained by various kinds of criteria such as event sets, biographical sets, …etc., as shown in Table 3.1.

Table 3.1: Statistic of DUC 2002 Data Set (Over 2002)

Category	Document Category
1	Single Natural disaster
2	Single event in any domain
3	Multiple distinct events of single type
4	Bibliographical information about a single individual

Three tasks were evaluated in DUC-2002: (1) Full automatic summarization of single newspaper/newswire document, (2) Full automatic summarization of several newspaper/newswire documents on single subject by producing extracts of documents and (3) Full automatic summarization of several newspaper/newswire documents on single subject by producing abstracts of documents. The most suitable data set chosen for our work is DUC 2002 as it refers to task3 (multi-document abstractive summarization on single subject) defined only for DUC 2002. This study will use DUC 2002 document sets for evaluation of the proposed framework.



Phase1

- ***Preliminary Study and Data Preparation***
- Problem formulation, literature review and selection of dataset

Phase2

- ***Extraction of predicate argument structure after semantic role labeling***
- This phase automatically identifies semantic representation (predicate argument structure(PAS) from text documents in order to generate a better summary.

Phase3

- ***Cross document relation identification using CBR approach***
- Develop a case base reasoning (CBR) classifier to automatically identify the existence of cross document relations (CST)between pair of PASs.

Phase4

- ***Semantic approach of multidocument abstractive summarization using SRL and CST***
- According to the number of relation types that each PAS holds a score is given to each PAS ,then the scored PASs were ordered using document NO and the sentence position NO in that document , lastly the selected higher scored PASs ordered to constitute final summary

Phase5

- ***Report writing***
- Write up the research report into chapters

Figure 3.1: The proposed operational research frame work

Data Pre-processingData Pre-processing is an important process in computational linguistics, particularly in text summarization. Since this work is concerned with text summarization; the input documents are usually in raw text format, which need to be pre-processed in order to transform them into appropriate representation that can be efficiently used in the experiments. In this study, we normally perform three text pre-processing steps i.e. sentence splitting, tokenization and (stop word elimination and word stemming).

a- Text segmentation

This step is the most fundamental part in natural language processing applications such as text summarization, information extraction, syntactic parsing, semantic role labeling and machine translation. The process of boundary detection and splitting the text into sentences is called sentence segmentation. Generally, a period (.), a question mark (?), or an exclamation mark (!) are the common signals that signify a sentence boundary (Mikheev et al. 2000.).

b- Tokenization

To do the task of tokenization, we use a simple program to split the sentences into distinct words by splitting them at whitespaces such as blanks, tabs and punctuation marks such as period, semicolon, comma, colon etc, are the main cues for splitting the text into tokens.

c- Stopwords

The Majority of text documents contain common or repetitive words which can make up 50% to 60% of a collection of documents text words. These words, usually called stopwords, are common to all domains and mostly belong to the word category type:-prepositions, conjunctions and articles. Examples of stop words are ‘the’, ‘a’, ‘and’,‘to’, ‘at’ and ‘on’. These words do not give much meaning but appears too frequentin a document. To avoid being considered as potential or important words, stopwords are normally removed from the texts. Besides that, eliminating stop words canalso speed up the system processing time, where fewer words need to be processed. The list of stop words is given in Appendix A.

d- Stemming

Stemming is a technique to find the root of words, so that the text processingis conducted on the roots and not on the original words. Through this process, theaffixes (prefixes and suffixes) in a word will be removed. For example, by using astemming algorithm, words such as ‘playing’, ‘played’ and ‘plays’ will be reducedto the root word ‘play’. Stemming proofs to be useful in information retrieval processlike in pattern or string matching where the existence of variance in word form canbe handled, i.e., allowing more terms to be unified in a document. The PorterStemming algorithm (Porter 1980) is most commonly used for the English language.We will also employ the Porter Stemmer for this purpose.

3.3.2 Phase 2: Extraction of predicate argument structure

In this phase we employ semantic role labeling which identifies semantic representation (predicate argument structures) from the document text automatically. The aim of semantic role labeling (SRL) is to determine the syntactic constituents/arguments for each predicate in a given sentence. We useSENNA toolkit to accomplish the role labeling ,(Aksoy et al. 2009)identifies the semantic roles of the arguments such as Agent, Patient etc., and the adjunctive arguments of the predicate such as Locative, Temporal etc.To accomplish this step we apply the following ordered steps:

(i) Data pre-processing
(ii) Extract semantic representation from text based on semantic role labeling technique
(iii) Computation of semantic similarity matrix

The details of each of these steps will be described later in Chapter 4.

3.3.3 Phase3 Cross document relation identification using CBR approach

In this phase we identify the CST (Cross Structure document relation Theory) between the PASs which are prepared in the previous phase.Since this study involves the automatic identification of cross document relations between PASs. Here we model a case based reasoning (CBR) classifier to classify the cross-document relations. There are four major steps in thecase base reasoning model, that is:
(i) Retrieve
(ii) Reuse
(iii) Revise
(iv) Retain

The details of each of these steps, corresponding to the cross-document relation classification will be described later in Chapter 5. For the experiment, we use the data set obtained from the CSTBank corpus (Radev 2002) which comprisesexamples of sentence pairs annotated with cross-document relations.. Five features will beextracted from each PASs pair to form its feature vector, namely; Synonyms Overlap(SO) in PAS, Type Lengthbased on length of (PAS) ,Noun-Phrase semantic similarity(NP) and Verb-Phrase semantic similarity(VP)and PAS to PAS semantic similarity(details are given in Chapter 5).To evaluate the performance of the proposed CBR classifier, we use four evaluation metrics which are commonly used for classification tasks, i.e. precision, recall, F-measure and accuracy (details are given in Section (4.5)). All these measures will be used to evaluate our test set and the performance will be compared against the selected benchmark methods used in this study.

3.3.4 Semantic approach of MDAS using SRL and CST

According to the number of relation types that each PAS holds(as we can find that each PAS can have relation with many other PASs)beside the use ofthe rules suggested in figure(6.1) a score is given to each PAS,the scored PASs were ordered using *document NO* and *sentence position NO*in that document , the over all summary is 20% ratio from all PASs , therefore we select the best 20% of highest scored PAS to PAS semantic similarity .

3.3.5 Writing up the research report

This phase will focus on putting up all the research components together to form the whole research and these components are introduction , literature review , research methodology , techniques used in research and their capabilities , experimental reports and general discussion about the whole work concluding the most important finding and drawing the future work plan into final thesis

3.4 Experimental evaluation measures

In this section, we describe the evaluation measures that will be carried out for each experimental phase in this research. In order to evaluate the system generated summary, standard evaluation measures need to be applied. The two standard evaluation metrics are usual used in text summarization evaluation, Recall-Oriented-Understudy-for-Gisting-Evaluation (ROUGE) (Lin 2004)and Pyramid (Nenkova et al. 2007).

Previous studies stated that ROUGE metric has been directed and used for the evaluation of extractive summaries. ROUGE score is the n-gram exact matches between system summary and human model summaries. This metric cannot capture semantically equivalent sentences(Khan et al. 2015b). The other evaluation metric, the pyramid metric is used for the evaluation of abstractive summaries. The obvious advantage of pyramid metric over the ROUGE is that it can capture different sentences in the summaries that uses different words but express similar meanings (Nenkova et al. 2007).

3.4.1 Pyramid measures

Pyramid is an evaluation method that includes two tasks: creating a pyramid by annotating model (human) summaries, and evaluating new summaries (peers) against a pyramid. The method requires multiple model summaries to evaluate a peer summary.

A pyramid is created by identifying Summary Content Units (SCUs), i.e., sets of contributors (text fragments) in the model summaries that express the same semantic content. SCUs that appear in all model summaries are given the highest weight, equal to the number of model summaries, and are placed at the top tier (layer) of the pyramid. SCUs appearing in one model summary are given the lowest weight of 1 and are placed at the bottom tier of the pyramid. The pyramid has tiers (layers) equal to the number of model summaries.

Pyramid score for a new peer summary is (the sum of weights of its peer SCUs) over (the average SCUs in the model summary), which is called mean coverage score or recall oriented pyramid score. Pyramid score is ranged from 0 to 1, the high score indicate that content of peer summary is high weighted. The three pyramid evaluation measures – Mean coverage score or Recall, Precision, and F-measure are given as follows:

Mean coverage score or recall oriented pyramid score for peer summary(Nenkova & Passonneau 2004) or candidate summary is computed as follows:

$$Mean\ Coverage\ Score = \frac{Total\ \ Peer\ SCUs\ Weight}{Average\ SCUs\ in\ the\ Model\ Summary} \qquad (3.1)$$

Where *SCUs* refers to the summary content units and their weights correspond to number of model (human) summaries they appeared in.

The precision for peer summary (Nenkova & Passonneau 2004) or candidate summary is computed as follows:

$$Mean\ Coverage\ Score = \frac{Number\ of\ Model\ SCUs\ expressed\ in\ Peer\ Summary}{Average\ SCUs\ in\ the\ Peer\ Summary} \qquad (3.2)$$

The F-measure for peer summary can be computed from equations (3.1) and (3.2) as follows:

$$F - Measure = \frac{2x\ Mean\ Coverage\ Score\ x\ Precision}{Mean\ Coverage\ Score + \ Precision} \qquad (3.3)$$

3.5 Summary

This chapter investigated the methodology which consists of the frame work which is divided to five phases each of them was discussed separately. Also the summary evaluation measures are discussed such as pyramid evaluation measures.

CHAPTERIV

A Model for Employing SRLTo Extract PAS

4.1 Introduction

In this chapter the study aims to represent a dataset from sentence-level form to predicate argument structure form, which is considered as a higher- level of abstraction. This new representation can be processed further in various applications such as text summarization and plagiarism detection. The SRL(Semantic Role labeling)techniqueis used to identify sentence constituents then ,the researcher implements a model to extract the predicate argument structure from the sentences that undergo SRL automatically , theresults are comparedto a manual predicate argument structure extraction, a good result has been achievedwith high precision and recall values.

4.2 Semantic role labeling (SRL)

Semantic role labeling (SRL) is a process to identify and label arguments in a text, as introduced in details in section 2.5.1. 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 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 pre-process 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 two groups :
core arguments (Arg) and adjunctive arguments (ArgM) as illustrated in Table 2.1. 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.

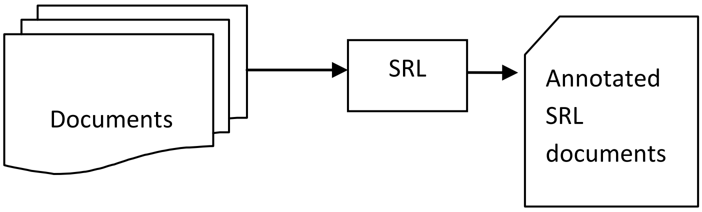
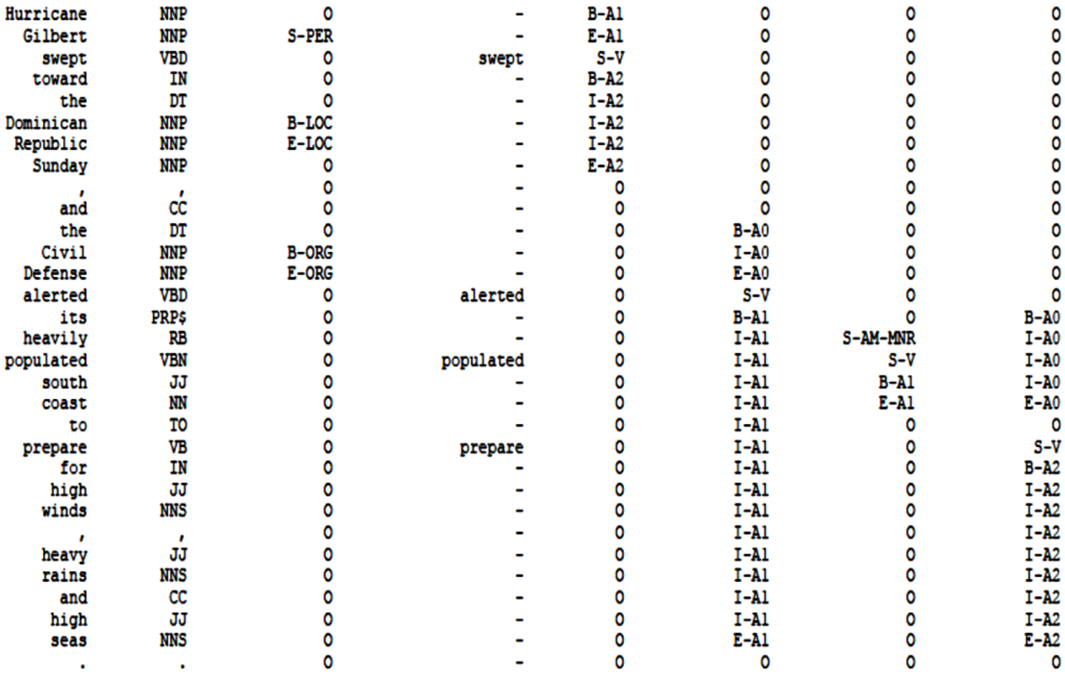


Figure 4.1 The general process of SRL



Hurricane	NNP	O	-	B-A1	O	O	O
Gilbert	NNP	S-PER	-	E-A1	O	O	O
swept	VBD	O	swept	S-V	O	O	O
toward	IN	O	-	B-A2	O	O	O
the	DT	O	-	I-A2	O	O	O
Dominican	NNP	B-LOC	-	I-A2	O	O	O
Republic	NNP	E-LOC	-	I-A2	O	O	O
Sunday	NNP	O	-	E-A2	O	O	O
,	,	O	-	O	O	O	O
and	CC	O	-	O	O	O	O
the	DT	O	-	O	B-A0	O	O
Civil	NNP	B-ORG	-	O	I-A0	O	O
Defense	NNP	E-ORG	-	O	E-A0	O	O
alerted	VBD	O	alerted	O	S-V	O	O
its	PRP$	O	-	O	B-A1	O	B-A0
heavily	RB	O	-	O	I-A1	S-AM-MNR	I-A0
populated	VBN	O	populated	O	I-A1	S-V	I-A0
south	JJ	O	-	O	I-A1	B-A1	I-A0
coast	NN	O	-	O	I-A1	E-A1	E-A0
to	TO	O	-	O	I-A1	O	O
prepare	VB	O	prepare	O	I-A1	O	S-V
for	IN	O	-	O	I-A1	O	B-A2
high	JJ	O	-	O	I-A1	O	I-A2
winds	NNS	O	-	O	I-A1	O	I-A2
,	,	O	-	O	I-A1	O	I-A2
heavy	JJ	O	-	O	I-A1	O	I-A2
rains	NNS	O	-	O	I-A1	O	I-A2
and	CC	O	-	O	I-A1	O	I-A2
high	JJ	O	-	O	I-A1	O	I-A2
seas	NNS	O	-	O	E-A1	O	E-A2
.	.	O	-	O	O	O	O

Figure 4.2 An example of an annotated sentence, in columns. Input consists of words (1st), PoS tags (2nd), named entities (3rd). The 4th column marks target verbs, and their propositions are found in remaining columns

Figure 4.2 shows an excerpt of a document after undergo SRL , the results are in columns where 1st column consists of the words tokens , 2nd column contain the Part of speech tagging , 3rd column is name entity recognition , 4th column is targeted verbs and remaining columns contains the role labeling for each targeted verb(predicate), as shown in figure 4.2 we have 4 targeted verb and this implies 4 columns for role labeling , each corresponds to one verb.

4.3 Predicate argument structures

The form of a predicate (verb) along with its Arguments is called predicate argument structure .In this study we consider two types of predicate argument structures (PAS) simple predicate argument structure and composite argument structure , the simple one is considered if we have one verb in a sentence and the composite one if we have more than one verb in the sentence , and we consider a PAS as a PAS if at least contains a verb and one other argument A0 or A1 .

Example

Consider the following sentence represented by composite predicate argument structures.

S*:Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas.*

The corresponding composite predicate argument structures PAS1 and PAS2 are obtained after applying semantic role labeling to sentence ***S***

PAS1: [A1: Hurricane Gilbert] [V: swept] [A2: toward the Dominican Republic Sunday].
PAS2: [A0: the Civil Defense][V: alerted][A1: its heavily populated south coast to prepare for high winds , heavy rains and high seas].

4.4 Experiment

The researcher intended to automate the extraction of PASs from each sentence therefore a useful system was implemented to extract the different PASs from each sentence , in order to accomplish this task a preprocess for the dataset was given such as removing the html tags and segmenting the text into separate sentences , then a SENNA toolkit was used to do the SRL , the SRL was given in terms of files , next the SRL files were entered to the extraction system one by one   and finally the system extracted the PASs from each sentence in an SRL file , the extracted PASs from each sentence are equal to the number of verbs in that sentence , which can be processed further for many other tasks such as summarization , categorization and classification.

4.5 Evaluation

We evaluate the results by comparing the system results against a manual one, a high precision and recall was given which asserted that the system model can be characterized as excellent as shown in figure 4.3.

$$Precision = \frac{SPP \cap HPP}{HPP}$$

$$Recall \quad = \frac{SPP \cap HPP}{SPP}$$

$$F\_Measure = \frac{2 \times Precision \times Recall}{Recall + precision}$$

Figure 4.3: Recall ,Precision and F_Measure formulasused for evaluation

Where *SPP* is System produced PASs and *HPP* is Human produced PASs .The average precision and recall for the tested DUC 2002 documents are shown in table (4.1).

Table 4.1 : Recall,Precision and F_measure for system

Precision	Recall	F_measure
0.905598	0.932293	0.918455

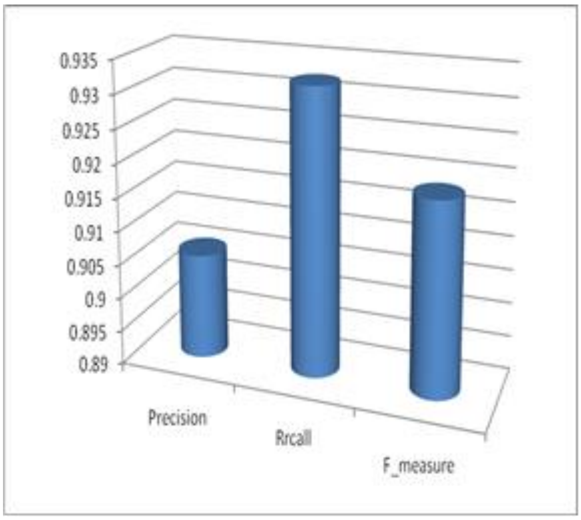


Figure 4.4: Recall ,Precision and F_measure

4.6 Summary

In this chapter SRL was employed for each sentence in a document in DUC 2002 dataset to extract predicate argument structure which is considered as semantic representation of sentence to be used further for other applications such as summarization, categorization and classification. To extract PASs from SRL, SENNA toolkit was used to employ SRL then it was entered to the designed system to extract PASs from those SRL files , as an evaluation method .the resulted PASs were compared to their manual peers and we got high precision , recall and F_measurewhich proves that our model can be characterized as excellent.

CHAPTERV

Cross-Document Relation Identification

Using Case Based Reasoning Approach

5.1 Introduction

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 2000) who explores that documents which are related to the same topic will contain semantically related textual units. Moreover he analysed and investigated the relationships that might exist between sentences across the related documents.

5.2 Overview of Approach

In this section, we provide an overview of our proposed approach for identifying cross-document relations from texts. We also describe the relations that will be used in particular for this study. There are four types of cross-document relations that have been considered here, namely Identity, Subsumption, Description and Overlap; since they cover most of the other relations in the CST model. For example, relations such as Historical Background and Contradiction are covered byDescription and Overlap. The explanation for each of the four cross-document relationsare provided in Table 5.1. All theserelations correspond to the relationships between two sentences (S1 and S2).

Table 5.1: Description of cross-document relations used in this study

Relation	Description
Identity	The same text appear in more than one location
Subsumption	S1 contains all information in S2, plus additional information not in S2
Description	S1 describes an entity mentioned in S2
Overlap (partial equivalence)	S1 provides facts X and Y while S2 provides facts X and Z

5.3 Related Works of CST based text summarization

The use of CST for multi document summarization is proposed to include and order the documents sentences based on their CST relation for summary generation and as mentioned earlier, previous works based on CST, regarded the CST types separately, 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 as shown in table (2.2), relations such as *Paraphrase, Modality* and *Attribution* share similar characteristic of information content with *Identity*.Consider the following example:

Example 1:

S1: RAI state TV reported that the pilot said the SOS was because of engine trouble.

S2: RAI state TV reported that the pilot said he was experiencing engine trouble.

Both sentences demonstrate an example of sentence pair that can represent *Identity, Paraphrase, Modality* and *Attribution* relations. The quality and amount of the information in both sentences are the same. Another example of sentence pair that can represent similar relations is shown in the following example:

Example 2:

S3: The crash put a hole in the 25th floor of the Pirelli building, and smoke was seen pouringfrom the opening.

S4: A small plane crashed into the 25th floor of a skyscraper in downtown Milan today.

Both sentences can be categorized as Elaboration and Follow-up. We can see from Example3 that Subsumption and Elaboration also shares some similar characteristics.

Example 3:

S5: The building houses government offices and is next to the city's central train station.

S6: The building houses the regional government offices, authorities said.

Thus, sentence pair connected as Subsumption can also be defined as Elaboration. However, sentence pair belongs to Elaboration in Example 2 cannot be defined as Subsumption. Here, Subsumption denotes S6 as the subset of S5, but as for Elaboration, S6 is not necessary a subset of S5. Therefore, we keep Subsumption and Elaboration as two different relations so that we can precisely perform the automated identification of discourse relation by using SVMs.

We redefined the definition of relations which was defined by(Radev 2004) as in Table 2, from CST by combining the relations types that resemble each other as described in Example 1, 2 and 3. Fulfilment by CST refers to sentence pair which asserts the occurrence of predicted event, where overlapped information present in both sentences.Therefore, we combined Fulfilment and Overlap as one type of relation. As for Change ofPerspective, Contradiction and Reader Profile, these relations generally refer to sentence pairs presenting different information regarding the same subject. Thus, we simply merged these relations as one group. We also combined Description and Historical Background, as both type of relations provide description (historical or present) of an event..

In this study some relations were combinedasfollows:

1- Combination of4 CST relations (modality , equivalence, attribution and identity) to be one group of CST regard it as "Identity".
2- Combination of 2 CST relations (Subsumption and Elaboration) to be one CST and regard it as "Subsumption".
3- Combination of 3 CST relations (Description, Historical ,Background) to be one CST and regard it as " Description ".

Each of the above groups of combination of relations has similar surface characteristics. The idea of combination is based on the study of (Adilah & Zahri 2012).The definition of each relation in accordance with the combination of relations is shown in table 5.2.

Table 5.2: The Proposed CST relations

Combined relations byCST	**Proposed Relations**	**proposed CST NO**	**Definitionof Proposed Relation**
Identity 1, Paraphrase 2, Modality 8,Attribution 9	Identity	1	Twotextspanshave thesame information content
Subsumption 4, Elaboration 13	Subsumption	4	S_1containsallinformationin S_2,plus otheradditionalinformation notin S_2
Overlap 18	Overlap	18	S_1providesfactsXand Ywhile S_2 providesfactsXand Z; X,Y, and Z should allbe non-trivial
Description 15, Historical Background 6	Description	15	S_1 giveshistoricalcontextordescribesan entitymentionedin S_2.
-	No Relations	0	No relation exits between $S_1$and S_2.

The combination of CST relations which are shown in table 5.2 refers to the similar characteristicsbetween each combined group of CST relations; this combinationsform new relations which are considered as a contribution in this study.

There are two reasonsfor using of CST relationships in this study. The first is that the study is conducted upon multidocument 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. The second reason is that the important information expressed in a sentence of a document is also expressed in the sentences of many related documents. For these two reasons weuse the (number of CST Relations that each PAS holds)in the calculation of the final score of a PAS.

5.4 Identification of CST Relations between PAS

We propose the use of CST for multidocument abstractive summarization to include and order the documents sentences based on their CST relations for summary generation. Previous works on multidocument summarization based on CST relations are employed for extractive summaries in which they regarded the CST types separately. Moreover they use plain text as dataset; where in our work which is mainly for abstractive multidocument summarization we concentrate on the PASs as representation of plain texts. We need to identify the CST relations among each pair of PASs. For the identification of these CST relations , we develop the CBR (Case Base Reasoning) classifier (Kumar et al. 2014) , we extract relevant features from each PAS pairs.

Earlier previous works based on CST, regarded the CST types separately(Lucía et al. 2011), 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(Zahri et al. 2015)(Zahri et al. 2015)(Zahri et al. 2015)(Zahri et al. 2015)(Zahri et al. 2015)(Zahri et al. 2015)(Zahri et al. 2015)).

In this study we propose five types of CST Relations which three types of them resulted as a combination of other types according to their similar surface characteristics as highlighted in table (5.2).

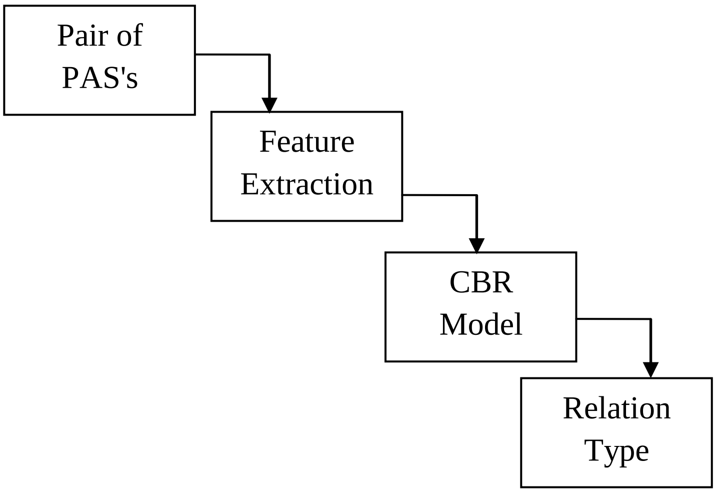


Figure 5.1 : Proposed method for CST identification

Each Pair of PASs will undergo five feature extraction ,these features are classified to specific relation type by using the CBR classifier model, the following section will discuss the proposed feature extraction.

5.5 Feature Extraction

Every PAS pair will be represented by its feature vector. The features areselected to adapt the related task to the problem of determining rhetorical status from texts. In this study, five features which compound of (deeper syntactic-level features) are unique to our cross-document relationship types are selected to represent each PAS pair (Maziero et al. 2011)(Maziero et al. 2011). 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. In the following section we provide the feature description for each of the mentioned features:

5.5.1 Synonyms Overlap in PAS

This feature represents the measure based on the number of overlapping words or synonyms of words based on wordNet between the two PASs(Kumar et al. 2014).

$$SynonymOverlapinPAS(SO) = \frac{\#\ commonwordsorsynonym(PAS1, PAS2)}{\#wordsorsynonym(PAS1) + \#wordsorsynonym(PAS2)} \quad Eq(5.1)$$

5.5.2 Type Length (based on Length(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(Pedersen et al. 1995).

$$Length(PAS) = \frac{NumberofwordsoccuringinthePAS}{NumberofwordsoccuringinthelongestPAS} Eq(5.2)$$

Type Length of PAS1 = 1 if Length(PAS1>Length(PAS2),
= -1 if Length (PAS1<Length(PAS2),
= 2 if Length(PAS1)=(Length(PAS2).

5.5.3 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 (Radev 2000)(Jiang & Conrath 1997).

$$Sem_{NP(}p_{i,}p_{j)=}\ Sem(NP_i\ ,NP_j) \qquad Eq(5.3)$$

5.5.4 Verb-Phrase (VP) Semantic Similarity

This feature determines semantic similarity between Verb-Phrase similarity in each pair of PAS using 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 (Radev 2000)(Jiang & Conrath 1997).

$$Sem_{VP(}p_{i,}p_{j)=}\ Sem(VP_i\ ,VP_j) \qquad Eq(5.4)$$

5.5.5 PAS to PAS Semantic Similarity

This feature computes the semantic similarity between pair of predicate argument structures. To compute the similarity between two PASs$(P_i, P_{\grave{j}})$ = 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:

$$sim_{arg}(P_i, P_{\grave{j}}) = sim(A0_i, A0_j) + sim(A1_i, A1_j) + sim(A2_i, A2_j)Eq(5.5)$$

$$sim_V(P_i, P_{\grave{j}}) = (sim(V_i\ , V_j\ ))Eq\ (5.6)$$

$$sim_{tmp}(P_i, P_{\grave{j}}) = (sim(Tmp_i\ , Tmp_j))Eq\ (5.7)$$

$$sim_{loc}(P_i, P_{\grave{j}}) = (sim(Loc_i\ , Loc_j\ ))Eq(5.8)$$

We combine Eq(5.5),Eq(5.6),Eq(5.7), and Eq(5.8) to give Eq(5.9).

$$sim\ (P_i, P_{\grave{j}}) = sim_{arg}(P_i, P_{\grave{j}}) + sim_V(P_i, P_{\grave{j}}) + sim_{tmp}(P_i, P_{\grave{j}}) + sim_{loc}(P_i, P_{\grave{j}})Eq\ (5.9)$$

And to apply Eq(5.9) we use Jiang's measure to compute the semantic distance to obtain semantic similarity for each part in Eq(5.9).

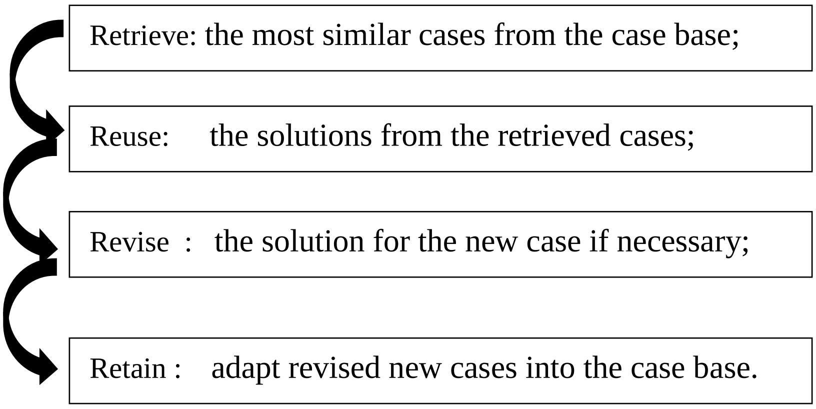
5.6 TheCase Base Reasoning CBR Approach

"Case Based Reasoning (CBR) is the usual name given to problem solving methods which make use of specific past experiences. It is a form of problem solving by analogy in which a new problem is solved by recognizing its similarity to a specific known problem, then transferring the solution of the known problem to the new one" (Bareiss et al. 1989). Using its memory of past experiences, CBR could be applied to solve various real world problems such as course timetabling, solving legal cases and classifying the disease of a patient(Kumar et al. 2013). Some of the prominent CBR applications which have been used extensively include systems such as Appliance Call Center automation at General Electric(Cheetham& Goebel 2007).

In our work, we also consider the task of identifying the cross-documentrelations between PAS pairs as a multi-class classification problem, whereby therelation can be classified to one of the following: Identity, Subsumption, Description,Overlap or No Relation.

The inclusion of No Relation is necessary as we cannotassume all sentence pairs to be related. Our method requires the adapttion of the standard CBR algorithm that is tailored to the classification task.

The general process of CBR consists of four major phases, namely *Retrieve*,*Reuse*, *Revise*, and *Retain* that links to a central repository called the case base(Aamodt& Plaza 1994); refer to Figure 5.2. When a new case (problem) is received, the CBR model will first retrieve the most similar cases from the case base (where previous solved cases are stored) and the solution from the retrieved cases will be reused for the new case. If no similar cases are found in the case base, the solution for the new case will be revised and retained into the case base as a new solved case. In the following subsection, we will show how the cases are represented in this study.



Retrieve: the most similar cases from the case base;

Reuse: the solutions from the retrieved cases;

Revise : the solution for the new case if necessary;

Retain : adapt revised new cases into the case base.

Figure 5.2: Phases in Case Base Reasoning process

5.6.1 Case Representation

Cases in a case base can represent the type of knowledge that it carries andsuch knowledge can be stored in various representational formats. The objective of acase based reasoning system is greatly influenced by what is stored in its case base.sets of attribute value pairs, i.e. the problem features and its solutions(Mántaras et al. 2005).

For example, in a medical CBR system (that is built to diagnose a patient), the case may be a set of symptoms (problem features) along with the diagnosis (solutions). Likewise, in this study, each case in our case base represents an example of PAS pair with its known cross-document relationship type. Specifically, every PAS pair will be labelled by the set of features that we described in Section 5.3;namely Synonyms Overlap, Type Length, (NP) Semantic Similarity, (VP) Semantic Similarityand PAS to PAS semanticsimilarity. Figure 5.3 shows how the cases are formed from the PAS pairs. Here, every PAS pair will be first pre-processed, i.e. by stop-word removal and word stemming. Then, feature extraction is performed on the PAS pair by computing the five abovementioned features using equations (5.1-5.5) as provided in Section 5.5. These feature values will then form the feature vector to represent eachcase. An example case representation in our case base is shown in Table 5.3. Theinput features f_1, f_2, f_3, f_4 and f_5 correspond to Synonyms Overlap(SO), Type Length, Noun-Phrase similarity (NP),Verb-Phrase similarity(VP) and PAS to PAS semantic similarity(PtoP) respectively, while the outputrepresents its solution (relationship type).

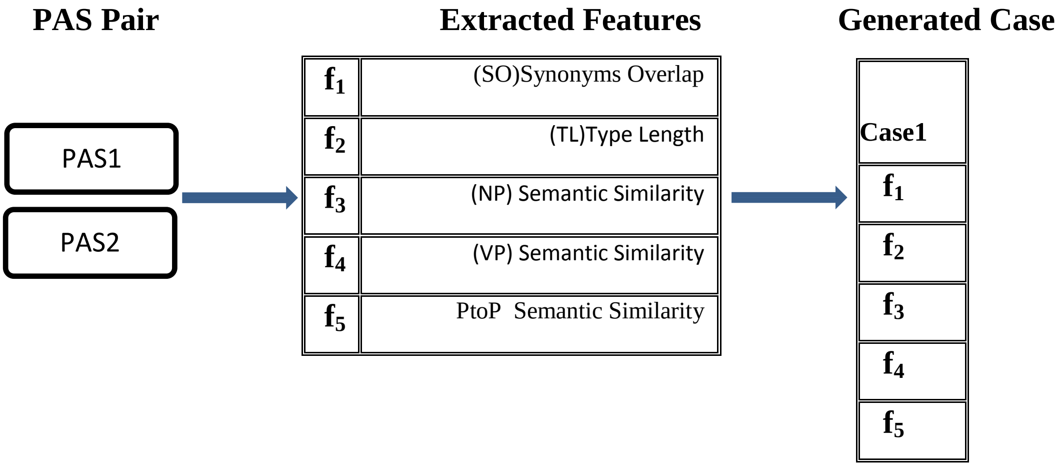


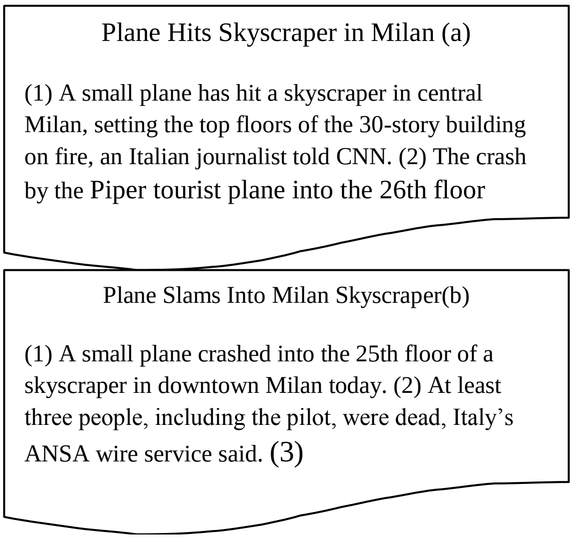
Figure 5.3: The process of generating cases from sentence pairs

Table 5.3: An example of case representation

Input features(f_i)						
Cases	f_1	f_2	f_3	f_4	f_5	**Output**
Case1	1	0.503505	0.665959	0.736002	0.166667	escription
Case2	1	0.495487	0.546225	0.108163	0.003907	ubsumption

5.6.2 The use of CBR in this study

n 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 5.4 , 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)(Radev 2002). For our study we will employ SRL for the dataset and extract PASs along with the features mentioned in section (5.5).Therefore we enrich the CBR Knowledge base with the pair of PASs features with respect to their annotated CST type. we train the CBR using the Enriched PAS and then we use the DUC 2002 PASs as testing ,accordingly the CST type will be identified that's the relation type as the CBR output.



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

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 ANSA wire service said. (3)

Figure 5.2 : Snapshot of CSTBank data set examples

5.6.3 The Identification of CST Relations between PASs using CBR

As the establishment of the representation of the cases and the store of them in the case base has been started, we can then model the case based reasoning (CBR) framework for the cross document relation identification. Our proposed CBR model is depicted in Figure 5.4. We firstperform pre-processing and feature extraction on the new PAS pair (that we aimto determine its relation). Once we have extracted all the features from the new PAS pair, we represent them as feature vector (input). To determine therelationship type for a new case, the CBR model will compare the feature vector ofthe new case with existing cases in the case base. Since each case (PAS pair) isrepresented as a vector, the similarity can be measured by using the cosine of thevectors. Cosine similarity gives a useful measure between two vectors and it is veryefficient to evaluate(Manning et al. 2008). Here, we define thecosine similarity measure to compute the similarity between two cases (PAS pairs) (*X*, *Y*);

$$\cos(X,Y) = \frac{\sum_{k=1}^{5} x_k \times y_k}{\sqrt{\sum_{k=1}^{5} x_k^2} \times \sqrt{\sum_{k=1}^{5} y_k^2}} \qquad (5.6)$$

where *cos*(*X*, *Y*) denotes the similarity between two cases (feature vectors *X* and *Y*),while ($x_k$and y_k) are the feature values corresponding to their *k*th feature, *k*={1,2,…,5}. The outcome of cosine measure is carefully bounded in [0, 1].

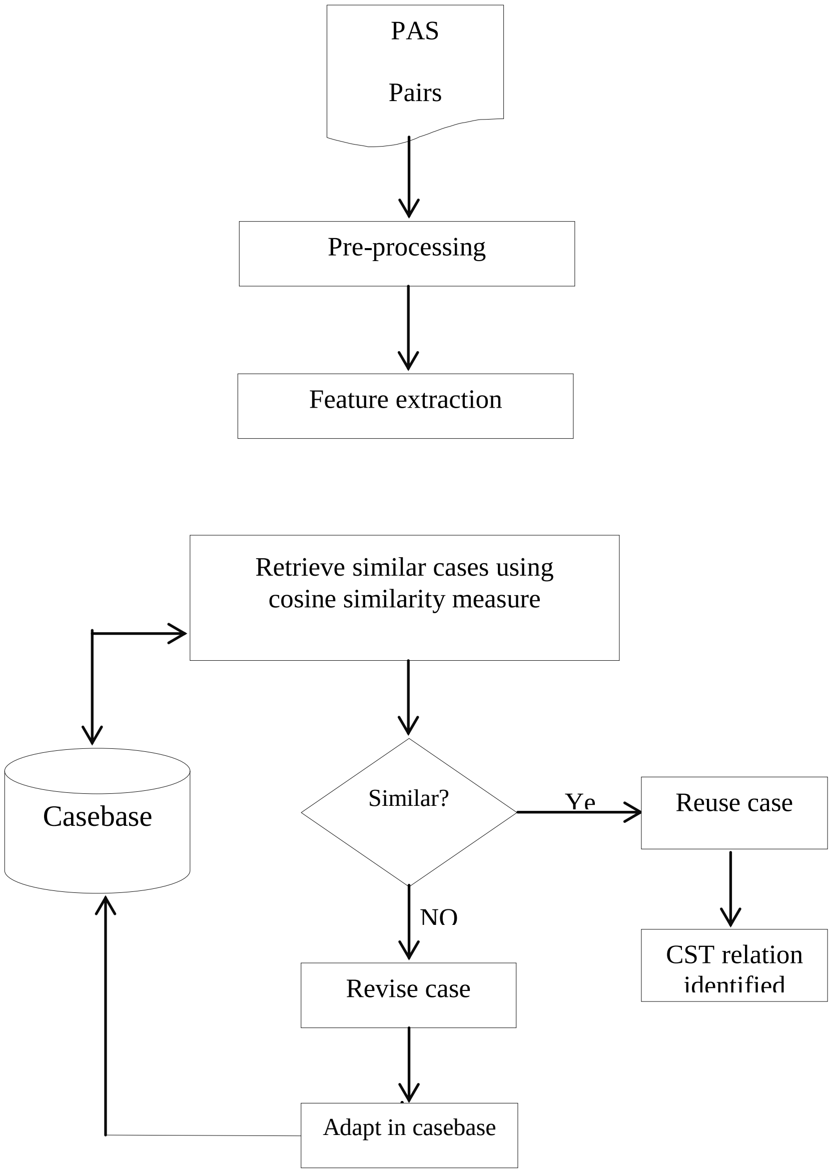


Figure 5.5: CBR approach for cross-document relation identification

An example of similarity computation using cosine similarity measure isgiven in Table 5.4, where a new case is being matched with Case 1 and Case 2 fromthe case base. Now using the similarity measure, all the similar cases from the casebase can be retrieved. If the similarity value of the new case is more than thepredefined threshold value 0.7, the model will reuse the solution (i.e. the relationshiptype of the existing case in the case-base). The threshold value was chosen based onexperimental observation. However if the similarity value is less than the thresholdvalue, the model will revise the new case as “No relation” type and retain the revisednew case into the case-base. Algorithm 5.1 below describes the CBR implementationwhile Figure 5.5 illustrates the overall process flow.

Table 5.4: An example of similarity computation between cases

Input Features	**NewCase**	**Case1**	**Case2**
Synonyms overlap	0.63	0.23	0.44
Type Length	0.51	0.36	0.34
NP semantic similarity	1	0	1
VP semantic similarity	0.42	0.27	0.55
PAS To PAS semantic similarity	0.47	0.16	0.36
Similarity with new case		0.68	0.97

Table 5.5: CBR approach for cross-document relation identification.

Steps	Main process	Process Detail
1.	Input PAS pairs	Take PAS pairs as input
2	Perform pre-processing	Perform stop word removal and word stemming
3	Features Extraction	Compute the features *F={SO, TL, NP, VP,PTP}* from each PAS pair to represent the cases.
4	Retrieve similar cases	Retrieve similar cases from the case base using the cosine similarity measure: equation (5.6).
5	Determine the cross-document relation of the new case	5.1 If the similarity value of the new case is more than the predefined threshold value, the model will reuse the solution i.e. the cross-documentrelation. *5.2* If the similarity value is less than the threshold value, the model will revise the new case solution as "No relation" and retain the revised case into the casebase.
6	Repeat steps	Repeat steps 4 to 5 for each sentence pair in the test set.

5.7 Experimental Setting

To experiment the implementation of the proposed CBR model for the task ofcross-document relation identification, we used the dataset obtained from CSTBank(Radev 2002) a corpus consisting clusters of English news articles annotated with cross-document relations. We collected 582 sentence pairs having the relation types Identity, Subsumption, Description and Overlap. We also manually selected 100 pairs of sentences that possess no cross-document relations. At first we perform text pre-processing on each of the sentence pairs. This involves stop word removal and word stemming. After pre-processing, we applied the semantic role labeling to have each sentence in a form of PAS then the features (as described in Section 5.5)were extracted. These features will then form the instances for the training set whereeach instance was represented as feature vector with its corresponding cross-document relationship type.

From the total 682 examples (including the 100 pairs with no relation), we selected 476 sentence pairs for training and 206 sentence pairs for testing. We implement our CBR model on the MATLAB platform. Both case-base (training set) and testing set were represented in matrix form where the rows represent the cases and the columns represent its input features. If there is more than one case with similarity value greater than the threshold, the program selects the highest value among them to be then retrieved. The retrieved case will then be reused to classify the tested-case. The CBR process (as shown in Figure 5.5) continues until all tested-cases have been classified.

5.8Experimental Results

To evaluate the CBR classifier we employ three evaluation measures Precision, Recall and F_measure,Table 5.5 and Figure 5.6 show the precision, recall, and F-measure of CBR classification.

The CBR and SVM classifiersweremodelledusingmatlab for the identification of the relation between each pair of PASs, the input to the both classifiers was(the five features plus their corresponding CST. The CST was coded as follows(1 for identity , 4 for subsumption , 15 for description, 18 for overlap and 0 for norelation).This experiment was done for melan dataset, the result is shown below:

Table 5.6: Precision, recall, and F-measure of CBR classification.

CST_Type	**Precision**	**Recall**	**F-Measure**
No Relation	0.8	0.121212	0.21052632
Identity	0.75	0.363636	0.48979592
Subsumption	0.4955	1	0.6626506
Description	0.84211	0.761905	0.8
Overlap	0.58621	0.523077	0.55284553

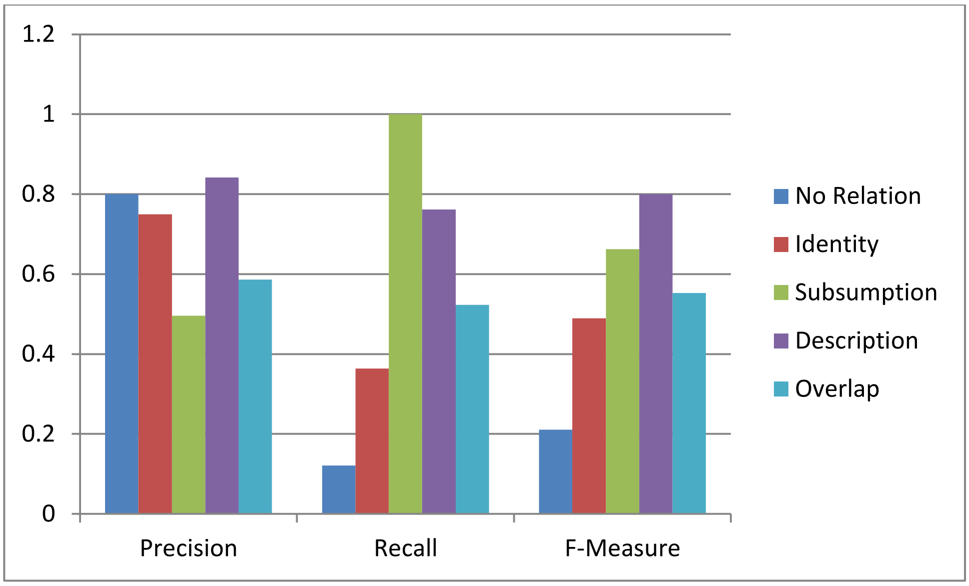


Figure 5.6: Performance of CBR classification.

Classifier Accuracy =0.60088

ent of our proposed method we compared it with SVM.We also tested the performance of support vector machine (SVM) using the same dataset. We chose the SVM as it is considered as a popular machine learning techniques commonly used for classification tasks (Kotsiantis 2007).

To evaluate the SVM classifier, we trained the data using the LibSVM tooldeveloped by(Chang & Lin 2011) on MATLAB. LibSVM is integratedsoftware which is extensively used for support vector classification and regressionfor solving multi-class classification problems. For the kernel selection, we chose theRBF kernel function as it gives better accuracy as stated by(Hsu et al. 2014) andithas several advantages over the other kernel functions. The SVM model bestparameters were chosen after applying 5-fold cross validation. Once the training iscompleted, the resulting classifier model is then tested with the test data to measureits performance. Table 5.6 and Figure 5.7 show the precision, recall, and F-measureof SVM classification.

Table 5.7: Precision, recall, and F-measure of SVM classification

CST_Type	Precision	Recall	F-Measure
No Relation	0.8	0.121212	0.21052632
Identity	0.775	0.939394	0.84931507
Subsumption	0.38596	0.8	0.52071006
Description	0.77273	0.404762	0.53125
Overlap	0.61538	0.492308	0.54700855

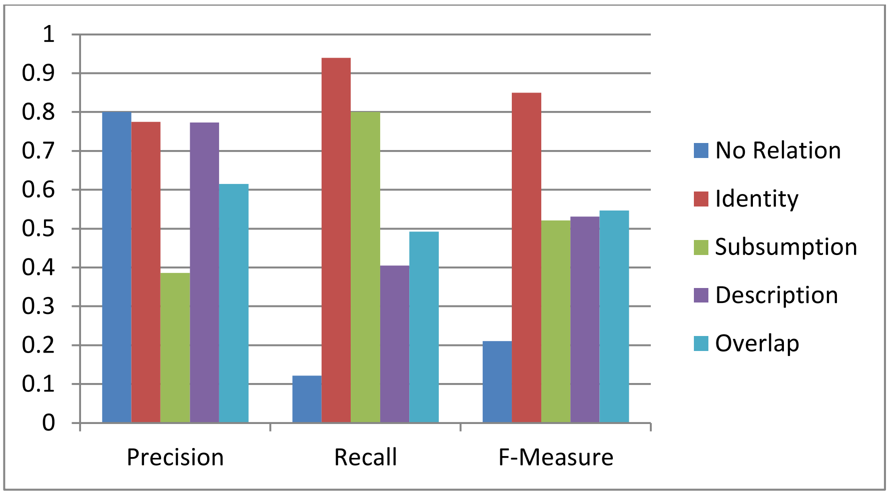


Figure 5.7: Performance of SVM classification.

ClassifierAccuracy =0.54386

As a comparison between both classifiers , figure 5.6 and figure 5.7 were combined in figure 5.8

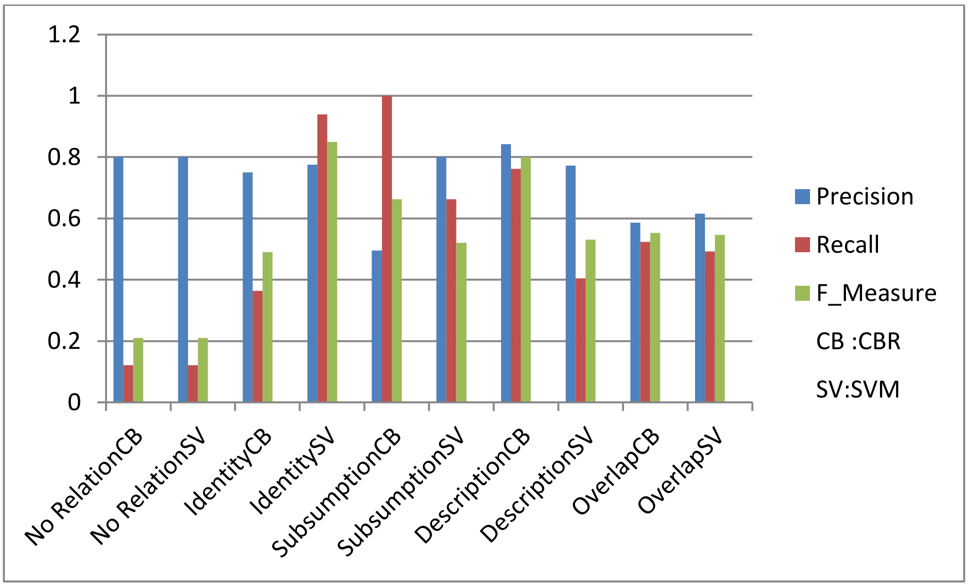


Figure 5.8: comparisonbetween CBR and SVM

5.9 Discussion

This section discuss the results shown in the previous section It isimportant to compare other machine learning techniques on the same data set, tosee if the performance of the technique being proposed is comparable or performsbetter than the other techniques. In this work for the automatic identification of cross-document relationships, we have compared the performance of our proposed CBR model with Support Vector Machine (SVM) which is also used for classification task. As presented in the previous section the performance of the CBR and SVM in table (5.6-5.7), we can observe the performance of each classifier in identifying the CST relationship types. Also the performance of each classifiers is given in which appear that our proposed CBR classifier performs better than SVM.

5.10 Summary

This chapter provides the study on cross-document relation identification between PASs in topically related documents. The need for automatic identification of cross-document relation is indeed necessary for tasks related to multi document analysis, for example, in multi document summarization tasks. Relying on manually annotated text for such tasks consumes a lot of time and resources. With the intention to have a system which can automatically identify the existence of cross-document relation between sentences in texts, we propose a supervised machine learning technique using the case based reasoning (CBR) model. We experimented the performance of the proposed method using the datasetobtained from CSTBank which comprises human annotated cross-document relations. We also describe in this work the implementation of a popular classification technique, i.e. support vector machine and compared its performance with our proposed method. Comparison between these techniques shows that the proposed CBR model yields better results.

CHAPTER VI

6 Generation of Multi document Abstractive Summarization

6.1 Introduction

In the previous chapter we have presented how to get the CST relation type between each pair of PASs , in this chapter we will discuss a method suggested by the researcher to getthe final summary, by combining these PASs according to specific rules .

6.2 Combine PASs based on CST

The researcher suggested to combine each pair of PASs that hold specific CST relationtype according to the rules illustrated in Figure 6.1.

Combine each pair of PAS's that have relations as the following

If CST is one of the following Do the decision:

Case CST between P1 And P2 is

Identity(1) : select P1

Case CST between P1 And P2 is

Subsumption (4) : select p1

Case CST between P1 And P2 is

Description(15) : select P1.

Case CST between P1 And P2 is

overlap : select P1 and p2

Case CST between P1 And P2 is

no relation(0) : select Both

Figure 6.1: CST Relations Combination Rules

6.2.1 PASs 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 an equal way in the summary, we suggest to assign the maximum similarity for the PAS with other PASs in the document in addition to the Number of CST relations which hold by the same PAS with other PASs in the document set, and this is considered as a final score for each PAS in order to get the best results. To achieve this, we will use equation Eq 6.1, Eq 6.2 and Eq 6.3.

$$PAS\_CST\_Score = \sum_{i=1}^{n} \#\ \text{Relation} \qquad (Eq\ 6.1)$$

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 (Eq 6.1) by adding a sort of fairness regarding the distributions of CST relations for each PAS with other PASs 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 PASs in document set as shown in (Eq 6.2).

$$PAS\_CST\_Score\_Ratio(p) = \frac{\sum_{i=1}^{n} \#\ \text{Relation}}{\sum_{j=1}^{N} \#\ \text{PAS} - 1} \qquad (Eq\ 6.2)$$

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

$$FinalPAS_{score} = PAS\_CST\_Score\_Ratio(p) + max(sim(P)) \qquad (Eq\ 6.3)$$

Where $sim\ \ (P)$ calculated by using$Eq(5.9)$then we select the maximum similarity which the current PAS has with other PASs in the document set in addition to $PAS\_CST\_Score\_Ratio(p)$ which calculated using (Eq 6.2).

6.2.2 Order PASsaccording to position in Source Text

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

6.2.3 Abstractive Summary

We combine each pair of PASs according to the rules suggested in figure 6.1 , then we select the best (highest (PAS To PAS) semantic similarity) as we can find that each PAS can have relation with many other PASs , the overall summary is 20% ratio from all PASs , therefore we select the best 20% of highest PAS to PAS semantic similarity.

6.3 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 PASs 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 PASs such as SO (Synonym Overlap), NP (Noun Phrase) similarity, VP(verb phrase) similarity , PAS to PAS similarity and PAS length , details are given in section (5.5) . These features are extracted and calculated by using the equations mentioned earlier in chapter 5 . Next we use a CBR classifier to identify the CST relation between each pair of PASs in the document , and then a combination between PASs is given to form the final summary .This combination is carried out according to rules suggested by the researcher which are given in section(6.2).

We employ both Pyramid and ROUGE evaluation metrics ,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.

6.4 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 (Over 2002). For each dataset, 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(Genest& Lapalme 2011) refers to the recent abstractive approach for multi-document summarization, Best automatic summarization system (Best) in DUC 2002, AS-SRL(Khan et al. 2015a) refers to semantic approach for multi-document abstractive summarization using semantic role labelingin DUC 2002, and the average of human model summaries (Models). For comparative evaluation, Table 4.2 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 4.7 visualizes the summarization results obtained with the proposed approach and other comparison models.

Table 6.1 : 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**

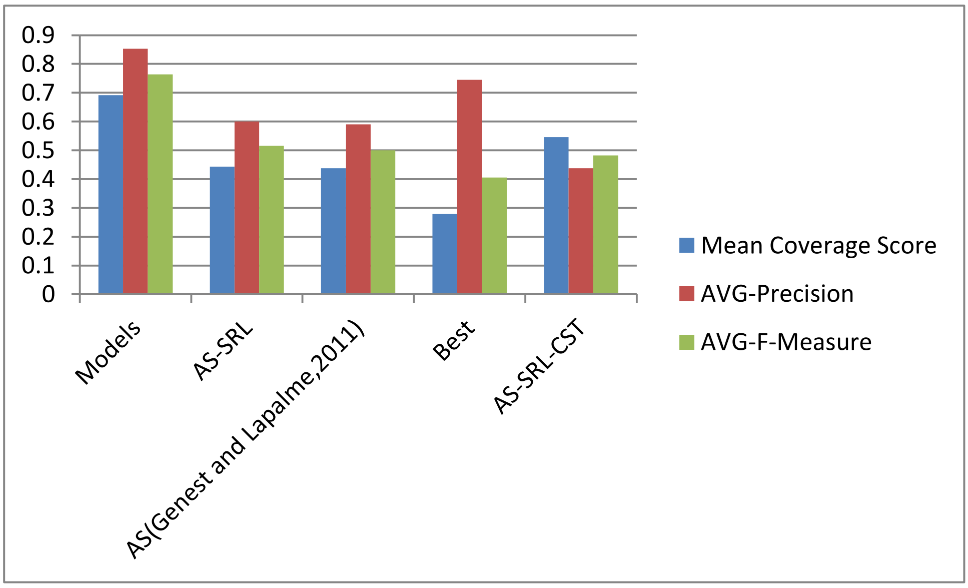


Figure 6.2: Comparison of summarization results based on mean coverage score, average precision and average F-measure

An explosion rocked the Royal Marines School of Music in a southeastern coastal town today, causing one building to collapse and killing eight people. The blast occurred at at 8:26 am in a lounge in the barracks near Deal, about 70 miles southeast of London.Scotland Yard said a forensic team from its anti-terrorist squad had been called in to help investigate.Firefighters used heavy lifting equipment and thermal cameras to search for those trapped in the debris, said Kent Fire Brigade spokesman Kevin Simmons.Kent police said 17 or 18 people were trapped.TheDefense Ministry said seven were missing.Ten doctors gave emergency treatment at the scene and 11 ambulances took the injured to two hospitals, the ambulance service said supplies to the barracks were cut as a precautionary measure, a spokesman said.Minnock's wife, Janet, said the roof of their house was torn off and all the back windows were shattered.GuyPlatts, who owns a bookstore in Deal, located 20 miles north of the English Channel port of Dover, said he heard a ``massive explosion.There are dozens of ambulances, police and fire brigade making their way there''.

Figure 6.3: Example of system generated summary

An explosion at 8:28 A.M. on 22 September destroyed a military barracks in Deal, a town

southeast of London in County Kent. The barracks was home to the Royal Marines School of Music. Ten military bandsmen were killed, and as many as 22 were injured, eight seriously. Neighbors expressed shock at the strength of the blast. No serious injuries to civilians were reported, but neighboring homes suffered heavy damage. The IRA claimed responsibility for the attack, linking it what they called a warlike speech delivered by Mrs. Thatcher during a recent visit to Northern Ireland. Investigators said the damage was illustrative of that caused by a bomb. The IRA has a record of attacking military installations, and this latest attack was the more damaging than any in more than seven years. The Prime Ministers of Britain and Ireland have denounced the

attack. Police are seeking three men with Irish accents who lived near the barracks for a short time.Private security firms serve the barracks at Deal and 29 other bases in Britain. Families of the victims and opposition party leaders have criticized the arrangement, blaming the attack on lax security because of the private guards.

Figure 6.4: Example of human produced summary

6.5 Discussion

It was observed from the results given in Table 6.1 , that mean-coverage score of the proposed approach (AS-SRL-CST) yields better summarization results than other comparison summarization models; and less in (AVG-Precision and average F-Measure), but still better than (Best-Model).

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.

6.6 Summary

This chapter explores the last construction of the summary usinga method suggested by the researcher to combine the PASs according to specific rules .More over a score is given to each PAS due to number of relation for each PAS with other PASs. And lastly 20% of the highest scored PASs were chosen to be included in the summary .also the summary has been evaluated by pyramid evaluation measures and compared against some models and gives a good results.

CHAPTER VII

7 CONCLUSIONS AND FUTURE WORK

7.1 Conclusion

Since the data on the Internetincrease day by day, people need a way to summarize all these data to a short version specially the data which is about specific event, we always find that the information is repeated in web sites and this is regarded as wasted time to go through these duplicated information, therefore people look forward to use text summarization to solve this duplication and to enable the readers to go through huge and short versions of non-redundant information .

Text summarization is divided into two disciplines extractive and abstractive which are defined as follows: the extractive one is to deal with the selection of an important terms from the original text and added them to the summary, here the text is reduced using the same words mentioned in the original text. On the other hand the abstractive approach require deeper analysis of text and the ability to generate new sentences, which provide an obvious advantage in improving the summary and reducing it is redundancy.

In this thesis 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 the source document to represent the text source semantically as Predicate Argument Structures (PASs).

Content selection for the summary is made by combining the PASs based on the Cross document Structure theory(CST) relations that each PAS has with other PASs, then according to number of relation types that each PAS holds we give a score to each PAS ,then we combine the PASs according to rules related to CST suggested by the researcher so as to reduce the redundancy, next the PASs were ranked using document No and the sentence position No in that document , lastly the selected higher scored PASsform the final summary.

7.2 Contributions

The main goal of this study is to introduce a Model of Abstractive Text Summarization for multidocuments by extracting Predicate Argument Structure out of sentences and to combine them using CST (Cross document relation Structure Theory) This above mentioned goal is satisfied by the following contribution.

- In feature extraction methods the researcher suggested to add a new feature and name it Synonym overlap (SO) which indicates the number of overlapping words or Synonym of words based on wordNet the lexical database for English between two PASs using the following formula:

$$SynonymOverlapinPAS(SOP) = \frac{\#\,commonwordsorsynonym(PAS1, PAS2)}{\#wordsorsynonym(PAS1) + \#wordsorsynonym(PAS2)} \quad Eq(5.1)$$

.

- In the CST (Cross document Structure Theory) The researcher suggests to combine some relations to be in one big relation due to similar characteristics between each combined group of CST relations,as illustrated in table 5.2 ,
- The researcher also suggests to combine each pair of PASs that hold specific CST relation type according to rules suggests by him.

.

7.3 Future Work

- Consider other domains where the application of multi-document summarization will be useful as summarizing academic journals and scientific papers
- Investigate how other cross –document relations can be identified from un-annotated text document by being able to identify the other relations.
- Consider Arabic languagefor abstractive text summarization.

Appendix A

STOPWORD LIST

A	B	C	D	E	F	H	I	M
a	be	can't	did	each	few	had	i	me
about	because	cannot	didn't		for	hadn't	i'd	more
above	been	could	do		from	has	i'll	most
after	before	couldn't	does		further	hasn't	i'm	mustn't
again	being		doesn't			have	i've	my
against	below		doing			haven't	if	myself
all	between		don't			having	in	
Am	both		down			he	into	
an	but		during			he'd	is	
and	by					he'll	isn't	
any	be					he's	it	
are						her	it's	
aren't						here	its	
as						here's	itself	
a						hers		
						herself		
						him		
						himself		
						his		
						how		
						how's		

N	O	S	T	U	V	W	Y
no	of	same	than	under	very	was	you
nor	off	shan't	that	until		wasn't	you'd
not	on	she	that's	up		we	you'll
	once	she'd	the			we'd	you're
	only	she'll	their			we'll	you've
	or	she's	theirs			we're	your
	other	should	them			we've	yours
	ought	shouldn't	themselves			were	yourself
	our	so	then			weren't	yourselves
	ours	some	there			what	you
	ourselves	such	there's			what's	you'd
	out		these			when	you'll
	over		they			when's	you're
	own		they'd			where	you've
			they'll			where's	your
			they're			which	yours
			they've			while	yourself
			this			who	yourselves
			those			who's	
			through			whom	
			to			why	
			too			why's	
						with	
						won't	

Appendix B

LIST OF PUBLICATIONS

#	Paper Title	Journal	Publication Date
1	Automatic Abstractive Summarization A SystematicLiteratureReview	Journal of Theoretical and Applied Information Technology August 2013 -- Vol. 54. No. 3 -- 2013	August 2013
2	A Model for Employing Semantic Role Labeling To Extract Predicate Argument Structure	International Journal of Computer Science Trends and Technology (IJCST) – Volume 4 Issue 5, Sep - Oct 2016	Oct 2016
3	A Model for Automatic Abstractive Multidocument Summarization	International Journal of Computer Science Trends and Technology (IJCST) in Nov 2018, Volume Number 6 Issue 6	Nov 2018,

References

Aamodt, A. & Plaza, E., 1994. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications*, 7(1), pp.39–59. Available at: http://iospress.metapress.com/index/316258107242JP65.pdf.

Adilah, N. & Zahri, H., 2012. Exploiting Discourse Relations between Sentences for Text Clustering. *Proceedings of the Workshop on Advances in Discourse Analysis and tis Computational Aspects(ADACA), Coling 2012*, (December 2012), pp.17–32.

Aksoy, C. et al., 2009. Semantic Argument Frequency-Based Multi- Document Summarization. In *2009 24th International Symposium on Computer and Information Science*. pp. 470–474.

Amini, M., Usunier, N. & Gallinari, P., 2005. Automatic Text Summarization based on Word-Clusters and Ranking Algorithms. *Advances in Information Retrieval*, pp.142–156. Available at: http://eprints.pascal-network.org/archive/00001070/.

Bareiss, R., Porter, B.W. & Murray, K.S., 1989. Supporting Start-to-Finish Development of Knowledge Bases. *Machine Learning*, 4(3), pp.259–283.

Barzilay, R. & Elhadad, M., 1997. Using Lexical Chains for Text Summarization. *Advances in automatic text summarization*, pp.111–121. Available at: http://portal.acm.org/citation.cfm?doid=1034678.1034760.

Barzilay, R. & McKeown, K.R., 2005. Sentence Fusion for Multidocument News Summarization. *Computational Linguistics*, 31(3), pp.297–328. Available at: http://www.mitpressjournals.org/doi/10.1162/089120105774321091.

Barzilay, R., McKeown, K.R. & Elhadad, M., 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* -. pp. 550–557. Available at: http://portal.acm.org/citation.cfm?doid=1034678.1034760.

Chang, C.-C. & Lin, C.-J., 2011. Libsvm. *ACM Transactions on Intelligent Systems and Technology*, 2(3), pp.1–27. Available at: http://dl.acm.org/citation.cfm?doid=1961189.1961199.

Cheetham, W. & Goebel, K., 2007. Appliance Call Center: A Successful Mixed-Initiative Case Study. *AI Magazine*, 28(2), pp.89–100. Available at: http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=25545228&site=ehost-live.

Edmundson, H.P., 1969. New Methods in Automatic Extracting. *Journal of the ACM*, 16(2), pp.264–285. Available at: http://portal.acm.org/citation.cfm?doid=321510.321519.

Gatt, A. & Reiter, E., 2009. SimpleNLG : A realisation engine for practical applications. *Proceedings of the ENLG '09 Proceedings of the 12th European Workshop on Natural Language Generation*, (March), pp.90–93.

Genest, P. & Lapalme, G., 2012. Fully Abstractive Approach to Guided Summarization. *50th Annual Meeting of the Association for Computational Linguistic*, (July), pp.354–358.

Genest, P.-E. & Lapalme, G., 2011. Framework for Abstractive Summarization using Text-to-Text Generation. *Workshop on Monolingual Text-To-Text Generation*, (June), pp.64–73. Available at: http://dl.acm.org/citation.cfm?id=2107687 [Accessed May 22, 2013].

Hovy, E. & Lin, C., 1996. Automated Text Summarization and the SUMMARIST System. *Proceedings of a workshop on held at Baltimore, Maryland October 13-15, 1998* -, p.197. Available at: http://portal.acm.org/citation.cfm?doid=1119089.1119121.

Hovy, E. & Lin, C., 1999. Automated text summarization in summarist. *Advances in Automatic Text Summarization*, pp.81–94. Available at: http://research.microsoft.com/en-us/um/people/cyl/download/papers/mit-book-paper-final-cyl.pdf.

Hsu, C., Chang, C. & Lin, C., 2014. Evaluating unsupervised and supervised image classification methods for mapping cotton root rot. , (April).

Jiang, J.J. & Conrath, D.W., 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. , (Rocling X), p.15. Available at: http://arxiv.org/abs/cmp-lg/9709008.

Johansson, R. & Persson, J., 2009. Text Categorization Using Predicate – Argument Structures Department of Computer Science. , I, pp.142–149.

Kasture, N.R. et al., 2014. A Survey on Methods of Abstractive Text Summarization. , (6).

Khan, A. & Salim, N., 2014. A review on abstractive summarization methods. *Journal of Theoretical and Applied Information Technology*, 59(1), pp.64–72.

Khan, A., Salim, N. & Jaya Kumar, Y., 2015a. A framework for multi-document abstractive summarization based on semantic role labelling. *Applied Soft Computing*, 30(July), pp.737–747. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1568494615001039.

Khan, A., Salim, N. & Jaya Kumar, Y., 2015b. A framework for multi-document abstractive summarization based on semantic role labelling. *Applied Soft Computing*, 30, pp.737–747. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1568494615001039.

Knight, K. & Marcu, D., 2000. Statistics-Based Summarization - Step One: Sentence Compression. *AAAI-00 - 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence*, pp.703–710. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.2394.

Kotsiantis, S.B., 2007. Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, 31, pp.249–268. Available at: http://books.google.com/books?hl=en&lr=&id=vLiTXDHr_sYC&oi=fnd&pg=PA3&dq=survey+machine+learning&ots=CVsyuwYHjo&sig=A6wYWvywU8XTc7Dzp8ZdKJaW7rc\npapers://5e3e5e59-48a2-47c1-b6b1-a778137d3ec1/Paper/p800\nhttp://www.informatica.si/PDF/31-3/11_Kotsiantis.

Kumar, Y.J. et al., 2013. Multi document summarization based on cross-document relation using voting technique. *Proceedings - 2013 International Conference on Computer, Electrical and Electronics Engineering: "Research Makes a Difference", ICCEEE 2013*, pp.609–614.

Kumar, Y.J. et al., 2014. Multi document summarization based on news components using fuzzy cross-document relations. *Applied Soft Computing*, 21, pp.265–279. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1568494614001598.

Kumar, Y.J., Salim, N. & Raza, B., 2012. Cross-document structural relationship identification using supervised machine learning. *Applied Soft Computing*, 12(10), pp.3124–3131. Available at: http://dx.doi.org/10.1016/j.asoc.2012.06.017.

Kupiec, J., Pedersen, J. & Chen, F., 1995. A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '95*. pp. 68–73. Available at: http://dl.acm.org/citation.cfm?id=215333 [Accessed May 27, 2013].

Lee, C.S., Jian, Z.W. & Huang, L.K., 2005. A fuzzy ontology and its application to news summarization. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 35(5), pp.859–880.

Lin, C.Y., 2004. Rouge: A package for automatic evaluation of summaries. *Proceedings of the workshop on text summarization branches out (WAS 2004)*, (1), pp.25–26.

Liu, F. & Liu, Y., 2009. From extractive to abstractive meeting summaries: can it be done by sentence compression? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. pp. 261–264. Available at: http://dl.acm.org/citation.cfm?id=1667583.1667664.

Lucía, M. et al., 2011. A Generative Approach for Multi-Document Summarization using the Noisy Channel Model. , pp.75–87.

Luhn, H.P., 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2(2), pp.159–165. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5392672 [Accessed May 27, 2013].

Malhotra, R., 2011. Software Effort Prediction using Statistical and Machine Learning Methods. , 2(1), pp.145–152.

Mani, I., 2001. *Automatic summarization*, John Benjamins Publishing.

Manning, C.D., Prabhakar, R. & Schutze, H., 2008. Introduction to Information Retrieval. *PhD Proposal*, 1, pp.139–161.

Mántaras, R.L.D.E. et al., 2005. Retrieval , reuse , revision , and retention in case- based reasoning. , 00, pp.1–2.

Màrquez, L. et al., 2008. Semantic Role Labeling: An Introduction to the Special Issue. *Computational Linguistics*, 34, pp.145–159. Available at: http://eprints.pascal-network.org/archive/00004514/.

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

Mikheev, A., Place, B. & Eh, E., Document Centered Approach to Text Normalization. , pp.136–143.

Moawad, I.F. & Aref, M., 2012. Semantic graph reduction approach for abstractive Text Summarization. In *Proceedings - ICCES 2012: 2012 International Conference on Computer Engineering and Systems*. Ieee, pp. 132–138. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6408498 [Accessed May 27, 2013].

Nenkova, A. & McKeown, K., 2012. A Survey of Text Summarization Techniques. In C. C. Aggarwal & C. Zhai, eds. *Mining Text Data*. Springer US, pp. 43–76. Available at: http://www.springerlink.com/index/10.1007/978-1-4614-3223-4.

Nenkova, A. & Passonneau, R., 2004. Evaluating content selection in summarization: The pyramid method. *Proceedings of HLT-NAACL*, 2004, pp.145–152. Available at: papers2://publication/uuid/DC675E84-0A45-48B7-A26C-F08B4B9398D3.

Nenkova, A., Passonneau, R. & McKeown, K., 2007. The Pyramid Method. *ACM Transactions on Speech and Language Processing*, 4(2), p.4–es. Available at: http://portal.acm.org/citation.cfm?doid=1233912.1233913 [Accessed March 6, 2013].

Over, P., 2002. Introduction to DUC-2002 : an Intrinsic Evaluation of Generic News Text Summarization Systems Document Understanding Conferences (DUC)….

Pedersen, J. et al., A Trainable Document. *Corpus*.

Porter, M.F., 1980. An algorithm for suffix stripping. *Program*, 14(3), pp.130–137.

Radev, D., 2000. A common theory of information fusion from multiple text sources. Step one: Cross document structure. *Proceedings of 1st SIGdial Workshop on Discourse and Dialogue*, pp.74–83.

Radev, D., Hovy, E. & McKeown, K., 2002. Introduction to the special issue on summarization. *Computational linguistics*, 28(4), pp.399–408. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21920798\nhttp://www.mitpressjournals.org/doi/abs/10.1162/089120102762671927.

Radev, D. & McKeown, K., 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3), pp.470–500. Available at: http://dl.acm.org/citation.cfm?id=972755.

Radev, D., Otterbacher, J. & Zhang, Z., 2004. CST Bank: A Corpus for the Study of Cross-document Structural Relationships. *Lrec*, pp.1783–1786. Available at: http://www.comp.nus.edu.sg/~rpnlpir/proceedings/lrec-2004/pdf/411.pdf.

Radev, D.R., 2002. Cross-document relationship classification for text summarization. *Computational Linguistics*.

Radev, D.R., 2004. LexRank : Graph-based Lexical Centrality as Salience in Text Summarization. , 22, pp.457–479.

Radev, D.R., Hovy, E. & McKeown, K., 2002. Introduction to the Special Issue on Summarization. *Computational Linguistics*, 28(4), pp.399–408. Available at: http://www.mitpressjournals.org/doi/10.1162/089120102762671927.

Suanmali, L., Salim, N. & Binwahlan, M.S., 2011. Fuzzy Genetic Semantic Based Text Summarization. *2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing*, pp.1184–1191. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6118856 [Accessed July 6, 2013].

Tanaka, H. et al., 2009. Syntax-driven sentence revision for broadcast news summarization. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation - UCNLG+Sum '09*. p. 39. Available at: http://www.aclweb.org/anthology/W/W09/W09-2808\nhttp://portal.acm.org/citation.cfm?doid=1708155.1708163.

Yao, B. & Li, S., 2010. ANMM4CBR: A case-based reasoning method for gene expression data classification. *Algorithms for Molecular Biology*, 5(1), pp.1–11.

Zahri, N.A.H., Fukumoto, F. & Suguru, M., 2015. E XPLOITING R HETORICAL R ELATIONS T O M ULTIPLE D OCUMENTS T EXT S UMMARIZATION. , 7(2), pp.1–22.

Zhang, Z., Blair-Goldensohn, S. & Radev, D., 2002. Towards CST-enhanced summarization. *Aaai/Iaai*, pp.439–445. Available at: http://www.aaai.org/Papers/AAAI/2002/AAAI02-067.pdf.