Review

A Review on Automatic Text Summarization Approaches

[1]Yogan Jaya Kumar, [1]Ong Sing Goh, [1]Halizah Basiron, [1]Ngo Hea Choon and [2]Puspalata C Suppiah

[1]*Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, 76100, Melaka, Malaysia*
[2]*Academy of Language Studies, Universiti Teknologi MARA (UiTM), 40450 Shah Alam, Malaysia*

Abstract: It has been more than 50 years since the initial investigation on automatic text summarization was started. Various techniques have been successfully used to extract the important contents from text document to represent document summary. In this study, we review some of the studies that have been conducted in this still-developing research area. It covers the basics of text summarization, the types of summarization, the methods that have been used and some areas in which text summarization has been applied. Furthermore, this paper also reviews the significant efforts which have been put in studies concerning sentence extraction, domain specific summarization and multi document summarization and provides the theoretical explanation and the fundamental concepts related to it. In addition, the advantages and limitations concerning the approaches commonly used for text summarization are also highlighted in this study.

Keywords: Automatic Text Summarization, Extractive Summarization, Domain Specific Summarization, Multi Document Summarization

Introduction

It has been more than 50 years since Luhn started his initial investigation on automatic text summarization (Luhn, 1958). Since then, various techniques have been successfully used to extract the important contents from text document to represent document summary (Gupta and Lehal, 2010; Nenkova and McKeown, 2011; Saggion and Poibeau, 2013). The aim of automatic text summarization is similar to the reason why we humans create summaries; i.e., to produce a shorter representation of the original text. Through these years, a number of researchers have defined the definition of summary from their own perspective. For instance, Sparck Jones defines a summary as a "reductive transformation of source text to summary text through content reduction by selection and generalization on what is important in the source" (Jones, 1999). Hovy defines a summary as "a text that is produced from one or more texts, that convey important information in the original text(s) and that is no longer than half of the original text (s) and usually significantly less than that" (Hovy, 2005).

Automatic text summarization systems can be categorized into several different types (Nenkova and McKeown, 2012; Saggion and Poibeau, 2013). The different dimensions of text summarization can be generally categorized based on its input type (single or multi document), purpose (generic, domain specific, or query-based) and output type (extractive or abstractive).

Single document summarization produces summary of single input document. On the other hand, multi document summarization produces summary of multiple input document. These multiple inputs are often documents discussing the same topic. Many of the early summarization systems dealt with single document summarization.

Generic summarization purpose is to summarize all texts regardless of its topic or domain; i.e., generic summaries make no assumptions about the domain of its source information and view all documents as homogenous texts. The majority of the work that has been done revolves around generic summarization (Nenkova and McKeown, 2011). There have also been developments of summarization systems which are centred upon various domain of interest. For example, summarizing finance articles, biomedical documents, weather news, terrorist events and many more (Radev and McKeown, 1998; Verma *et al.*, 2007; Wu and Liu, 2003). Often, this type of summarization requires domain specific knowledge bases to assist its sentence selection process. Query-based summary contains only information which are queried by the user. The queries are typically natural language questions or keywords that are related to a particular subject. For instance, snippets produced by search engines is an example of query-based application (Nenkova and McKeown, 2011).

Extractive summaries or extracts are produced by identifying important sentences which are directly selected from the document. Most of the summarization systems that have been developed are for extractive type summaries (Aliguliyev, 2009; Ko and Seo, 2008). In abstractive summarization, the selected document sentences are combined coherently and compressed to exclude unimportant sections of the sentences (Ganesan *et al.*, 2010; Khan *et al.*, 2015).

In this study, the study will focus on extractive based text summarization and will primarily review approaches concerning sentence extraction, domain specific summarization and multi document summarization methods.

The following section presents the details on approaches to sentence extraction. Then, the discussion on domain specific summarization is given. Following that, the discussions on multi document summarization approaches are presented and finally the paper ends with conclusion.

Approaches to Sentence Extraction

The key concept of extractive summarization is to identify and extract important document sentences and put them together as a summary; i.e., the generated summary is a collection of original sentences. There are several approaches to sentence extraction. The following subsections will describe three approaches, namely, frequency based approach, feature based approach and machine learning based approach.

Frequency Based Approach

As we discussed in the introduction section; in the early work on text summarization, which was pioneered by Luhn, it was assumed that important words in document will be repeated many times compared to the other words in the document (Luhn, 1958). Thus Luhn proposed to indicate the importance of sentences in document by using word frequency. Since then, many of the summarization systems use frequency based approaches in their sentence extraction process (Klassen, 2012). Two techniques that use frequency as a basic form of measure in text summarization are: word probability and term frequency-inverse document frequency.

A. Word Probability

It was assumed that one of the simplest ways of using frequency is by taking the raw frequency of a word i.e., by simply counting each word occurrence in the document. However, this measure can be greatly influenced by the document length. One way to make adjustment for the document length is by computing the word probability. The probability $f(w)$ of a word w is given by Equation 1:

$$f(w) = \frac{n(w)}{N} \tag{1}$$

Where:
$n(w)$ = The frequency count of the word w in the document
N = The total number of words in the document

The findings from the analysis carried by Nenkova *et al.* (2006) on human-written summaries indicate that people tend to use word frequency to determine the key topics of a document. SumBasic (Nenkova and Vanderwende, 2005) is an example of summarization system that exploits word probability to create summaries. The SumBasic system first computes the word probability (as given in Equation 1) from the input document. Then for each sentence Sj, it computes the sentence weight as a function of word probability (Equation 2):

$$weight(S_j) = \frac{\sum_{w \in S_j} f(w)}{\left|\left\{w \middle| w \in S_j\right\}\right|} \tag{2}$$

Based on the sentence weight, it then picks the best scoring sentences. Despite its simplicity (using only word probability), the SumBasic system was able to perform well in the Document Understanding Conference (DUC) 2004.

B. Term Frequency–Inverse Document Frequency

Term frequency-inverse document frequency (*tf-idf*) has been traditionally used in information retrieval to deal with frequent occurring terms or words in a corpus consisting related documents (Jurafsky and Martin, 2009). Its purpose was to address the following question: Are all content words that frequently appear in documents are equally important? For instance, a collection of news articles reporting on earthquake disaster will obviously contain the word 'earthquake' in all documents.

Thus the idea of *tf-idf* is to reduce the weightage of frequent occurring words by comparing its proportional frequency in the document collection. This property has made the *tf-idf* to be one of the universally used terminologies in extractive summarization (Filatova and Hatzivassiloglou, 2004; Fung and Ngai, 2006; Galley, 2006; Hovy and Lin, 1998). Here, the term frequency (*tf*) is defined as:

$$tf_{i,j} = \frac{n_{i,j}}{\sum n_j} \tag{3}$$

Where:
$n_{i,j}$ represents the frequency count of the word i in document j.

Each word is then divided or normalized by the total number of the words in document *j*. This term weight computation is similar to the word probability computation given in Equation 1. Next, the inverse document frequency (*idf*) (Jones, 1988) of a word *i* is computed:

$$idf_i = log \frac{|D|}{|\{d \mid t_i \in d\}|} \qquad (4)$$

where, the total number of documents in the corpus is divided by the number of documents that contain the word *i*. Based on Equation 3 and 4, the *tf-idf* of word *i* in document *j* is computed:

$$tf - idf_i = tf_{i,j} \times idf_j \qquad (5)$$

Feature Based Approach

One of the natural way to determine the importance of a sentence is to identify the features that reflects the relevance of that sentence. Edmundson (1969) defined three features deemed indicative to sentence relevance i.e., sentence position, presence of title word and cue words. For example, the beginning sentences in a document usually describes the main information concerning the document. Therefore, selecting sentences based on its position could be a reasonable strategy. The following features are commonly used to determine sentence relevance (Gupta and Lehal, 2010).

Title/Headline Word

Title words appearing in a sentence could suggest that the sentence contains important information.

Sentence Position

The beginning sentences in a document usually describes the main information concerning the document.

Sentence Length

Sentences which are too short may contain less information and long sentences are not appropriate to represent summary.

Term Weight

Words or terms which have high occurrence within a document is used to determine the importance of a sentence.

Proper Noun

Proper noun and named entities such as person, organization and location mentioned in a sentence are considered to be carrying important information.

Figure 1 depicts the general model of a feature based summarizer. The scores for each feature are computed and combined for sentence scoring. Prior to sentence scoring, these features are given weights to determine its level of importance. In this case, feature weighting will be applied to determine the weights associated to each feature and the sentence score is then computed using the linear combination of each feature score multiplied by its corresponding weight:

$$Score = \sum_{i=1}^{n} w_i \times f_i \qquad (6)$$

Where:
w_i = The weight of feature *i*
f_i = The score of feature *i*

Binwahlan *et al*. (2009) proposed a text summarization model based on Particle Swarm Optimization (PSO) to determine the feature weights. Bossard and Rodrigues (2011) used genetic algorithm to approximate the best weight combination for their multi document summarizer. Differential evolution algorithm has also been used to scale the relevance of feature weights (Abuobieda *et al*., 2013a). Investigation on the effect of different feature combination was carried by Hariharan (2010), where it was found that better results were obtained by combining term frequency weight with position and node weight.

In later works, the incorporation of fuzzy rules was studied by Suanmali *et al*. (2009; 2011) for scoring sentences. For instance, one of their constructed rules states "*if (NoWordInTitle is VH) and (SentenceLength is H) and (TermWeight is VH) and (SentencePosition is H) and (SentenceSimilarity is VH) and (ProperNoun is H) and (ThematicWord is VH) and (NumbericalData is H) then (Sentence is important)*". Their experimental finding (tested on the DUC 2002 data set) showed that the fuzzy logic based method could outperform a general statistical method. A recent study also supports the advantages of using fuzzy reasoning to determine the importance of a sentence (Babar and Patil, 2015).

Machine Learning Approach

Machine Learning (ML) approach can be applied if we have a set of training document and their corresponding summary extracts (Neto *et al*., 2002). The objective of machine learning can be closely related to a classification problem, i.e., to learn from a training model in order to determine the appropriate class where an element belongs to. In the case of text summarization, the training model consists of sentences labelled as "summary sentence" if they belong to the reference summary, or as "non-summary sentence" otherwise. Sentences are usually represented as feature vectors.
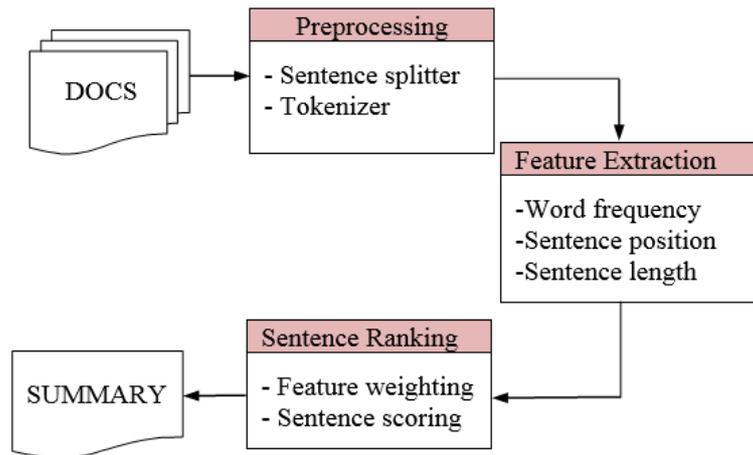
Fig. 1. A feature based summarization model

After learning from the collection of documents and its summaries, the trained model will be able to identify potential summary sentences when a new document is given to the system. Next we will discuss some related works on machine learning methods.

A. Naive Bayes

One of the early works that incorporated machine learning was the study done by Kupiec *et al.* (1995). They used a Naive Bayes classifier for learning from the data (corpus of document/summary pairs). Their method uses the features that were derived from Edmundson (1969), where the features were independent of each other. Given a sentence *s*, the probability of it being chosen to be included in the summary is:

$$P(s \in S \mid F_1, F_2, ..., F_n) = \frac{\prod_{i=1}^{n} P(F_i \mid s \in S) \cdot P(s \in S)}{\prod_{i=1}^{n} P(F_i)} \quad (7)$$

Where:
F_1, F_2, …, F_n are the sentence features (assuming the features are independent of each other) for the classification and *S* is the summary to be generated.

Each sentence is then scored according to Equation 7 and ranked for summary selection.

The Naive Bayes classifier was also used in later works but with richer features. Aone *et al.* (1999) include feature like *tf-idf* using noun words and named entities, where they used the corpus consisting of news documents for their experiments. Another extensive investigation using the similar framework was carried by Neto *et al.* (2002). The authors employ a large variety of features, including both statistical and linguistic features. Their method which uses the Naive Bayes classifier

significantly outperformed all the baseline methods. From their findings, they also reported that the choice of classifier can strongly influence the performance of the summarizer.

B. Neural Network

Some researchers have utilized the advantages of neural network learning capabilities to learn summary sentence attributes. Kaikhah (2004) used a three layered Feed-forward network model to learn the patterns in summary sentences (Fig. 2). Seven features were extracted from their input sentences. Once the network learns the features that best represent summary sentence, feature fusion was performed by removing and combining certain features. The pruned network model is then applied to determine the summary sentences.

In another related work, a single document summarization system called NetSum was developed (at Microsoft Research Department) by Svore *et al.* (2007). The system was built to generated summaries using a neural network model. First, the training set (articles collected from CNN.com) is used to train the network model. The trained model is then used to rank new sentences. The NetSum system uses the RankNet algorithm (Burges *et al.*, 2005) to perform sentence ranking. Based on the evaluation, it was found that NetSum achieved better results with statistically significant improvements compared to the baseline.

There are also other machine learning methods that has been recently used for text summarization. Hannah and Mukherjee (2014) proposed a trainable summarizer for classifying important sentences. The authors used a decision tree model which was trained to classify sentences as interesting sentence and not interesting sentence. The results of their approach was able to outperform the baseline approach results.
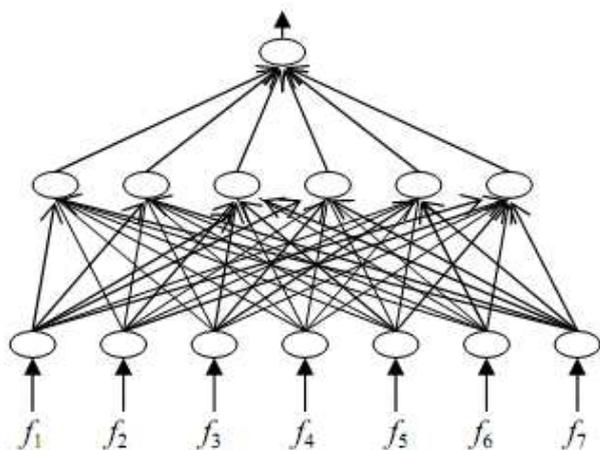
Fig. 2. Feed-forward network model after training (Kaikhah, 2004)

Domain Specific Summarization

Much of the work we reviewed in the previous sections involved generic summarization whereby the relevance of a summary is decided just based on the input document without relating to its domain or the user needs (Nenkova and McKeown, 2011). For example, inputs such as medical documents, news documents or emails; have special structures or unique characteristics which should be taken into account by the summarizer to produce more accurate information. Next, we will review some of the works concerning domain specific text summarization.

Medical Summarization

The study on automatic summarization was found to be very useful to the medical field. Summarization can help doctors to obtain relevant information about a particular disease or information from the patient records (Becher *et al.*, 2002). It will also be beneficial to patients or users whom turn online to find information pertinent to their health problems (Kaicker *et al.*, 2010). Furthermore, there are extensive resources that provide access to medical information and medical-related databases. For instance, there are over 20 million articles in MEDLINE; a biomedical database. Summarization is thus essential in such condition to treat the problem of information overload.

An early summarization system that has been built for medical knowledge is the Centrifuser (Elhadad *et al.*, 2005; Kan *et al.*, 2001). The Centrifuser is a summarizer that helps consumers by producing query-driven summaries in their search for healthcare information. It represents document topics by a tree data structure and perform query mapping from the topic trees to retrieve relevant sentences. Another medical summarizer, proposed by Fiszman *et al.* (2009), was built to generate summaries based on semantic abstraction to assist

physicians find the most salient information in MEDLINE citations for some specified diseases.

There are also researchers who utilize the background knowledge (i.e., ontology) for medical summarization. Ontology can be used to describe domain–related information. Using ontology, information can be related to each other through the common characteristics of a domain (Khelif *et al.*, 2007).

One example is the utilization UMLS, a medical ontology, which is used to summarize biomedical articles (Verma *et al.*, 2007). UMLS was used to match words in sentences that contains similar concepts in it. Likewise, Kogilavani and Balasubramanie (2009) have employed UMLS to expand user's natural language queries with synonyms and semantically related concepts. Ontology has also been used by Naderi and Witte (2010) in biomedical research area to summarize protein mutation impact information. They populated their ontology with protein mutation impact information and then used it to generate query based summaries.

News Summarization

Early work on news summarization can be dated back to 1990s when SUMMONS summarizer was created (McKeown and Radev, 1995). SUMMONS was designed for summarizing single events (news articles related to terrorist events). It was built using a template-driven message understanding system, MUC-4 (Sundheim, 1992). The system first processes the full text and fills the template slots before synthesizing the summary from the extracted information.

Similar to the SUMMONS system is a system called RIPTIDES (White *et al.*, 2001). It incorporates information extraction to support summarization. They use natural disaster scenario templates for each text and provide them as input to the summarization system. The summarizer first merges the templates into event oriented structure and then the importance scores are assigned to each slot/sentence to select the summary sentences.

Newsblaster (McKeown *et al.*, 2002), was developed to summarize online news articles. The summarizer uses MultiGen (Barzilay *et al.*, 1999; McKeown *et al.*, 1999), which identifies common sentences from news article using machine learning together with statistical techniques (Hatzivassiloglou *et al.*, 1999; 2001). Summaries are then produced by analyzing and fusing together the sentences.

In later work, Li *et al.* (2010) proposed Ontology-enriched Multi-Document Summarization (OMS) system to generate query-relevant summary applied to disaster management; for natural calamities related news and reports. OMS relates sentences onto a domain-specific ontology. Node on the ontology will then be matched based on user query and the sentences attached to that particular node will be extracted to form summary.
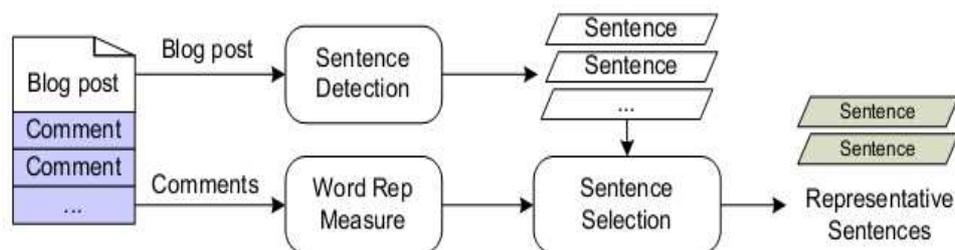
Fig. 3. Comment-oriented blog summarization (Hu *et al.*, 2007)

Another concept called fuzzy ontology was studied by Lee *et al.* (2005) to develop weather news summarization. Fuzzy ontology was found to be more suitable to treat domains with uncertainty.

From the understanding of news structure, the utility of sub-events in news topic were investigated by Daniel *et al.* (2003) in order to capture essential information to produce better summaries. Their study involves experiments carried out on Gulf Air crash. In their experiment, human judges were asked to determine the sentences related to the predefined sub-events comprising the topic. Then summaries were created using selection algorithms. Their findings showed that the utilization of sub-events can improve the performance and suggest that future efforts should be directed towards enhancing automatic clustering of sub-events. In another related work, Kumar *et al.* (2014) exploited news structure by incorporating the contextual information such as 'who', 'what' and 'where' in the sentence selection process. Contextual information was able to significantly improve news summarization.

Email/Blog Summarization

There have also been studies on email and blog summarization reported in literature. In early research on email summarization, Nenkova and Bagga (2004) came up with a system to generate summaries from email threads. They produce short "overview summaries" by extracting sentences only from the thread root message and its immediate follow-ups. To extract sentences from the root message, they find sentence that has the largest overlap nouns or verbs with the subject of the email. Similarly, to select sentence from follow-up emails, the largest overlap of nouns and verbs between the root email and the follow-up emails were computed.

Newman and Blitzer (2003) also address the problem of summarizing email threads. First, all the messages are clustered into group messages. Sentences in each group are scored using several features. Then from each group, summaries are extracted. In another related work, Rambow *et al.* (2004) used email specific features and rules to extract sentences from emails. The features that they used take into account the structure of the email thread.

For research in blog summarization, the main context of the blogs is usually the writer's opinion. Zhou and Hovy (2006) proposed a summarization approach which was inspired by the work by Marcu (1999), who produces summary extracts using (abstract, text) tuples. Starting from the blog entry, they continuously remove sentences that are not related to the story (linked articles), while keeping sentences with maximal semantic similarity with the linked articles.

In later work, Hu *et al.* (2007) argued that comments from blog readers does change the understanding about the blog post. The authors use the words from the blog's comments to extract sentences. They integrate several word representative measures to weight the words appearing in the comments and perform sentence selection based on the representativeness of its contained words. Figure 3 show the architecture of their blog summarization model.

Apart from personal blogs, summarization for legal blog entries has also been studied. Conrad *et al.* (2009) proposed a query-based summarization approach which is specific to legal blogs. The task carried was based on the Text Analysis Conference (TAC) 2008 task. Using the retrieved documents from the Blog Search Engine (www.blogsearchengine.com), they first filter the sentences that do not match the query questions (questions related to topics from the legal domain). Then they apply the FastSum (Schilder and Kondadadi, 2008); a summarization system which have been previously used to produce sentiment summaries (Schilder *et al.*, 2008a; 2008b), to extract summaries from the retrieved blogs.

Multi Document Summarization

Concerns have been raised in past regarding the size of input documents which is required to be summarized. Since information can be collected from multiple sources, condensing these information is considered essential. Various types of multi document summarization methods have been developed by researchers (Nenkova and McKeown, 2012; Saggion and Poibeau, 2013). In this section we will focus on two popular methods i.e., cluster based method and graph based method (Gupta and Lehal, 2010; Haque *et al.*, 2013). Besides these two methods, we also review some of the related works, using discourse based method, which received much attention in recent years. For each of these methods, its primary concept will be explained.

Cluster Based Method

Clustering refers to the grouping of similar instances into their clusters. In our case, these instances are the sentences. This can be done by computing the similarity between sentences and the sentences which are highly similar to each other are grouped into the same cluster. Different clusters may represent different subtopics. High scoring sentences from each cluster are then put together to form summary. This process is depicted in Fig. 4.

Radev *et al.* (2004) pioneered the use of cluster centroids for their multi-document summarizer, MEAD. Centroids are the top ranking *tf-idf* that represents the cluster. These cluster centroids are then used to identify the sentences in each cluster that are most similar to the centroid. The cosine similarity measure was used for this purpose. As a result, the summarizer generates sentence which are most relevant to each cluster.

Taking the benefit of clustering approach, efforts have been put into making the overall text summarization process more effective. One that is worth to be mentioned here is determining the optimal number of clusters, where Xia *et al.* (2011) adopted the co-clustering theory to find optimal clusters. They determine the weights of sentences and terms based on the sentence-term co-occurrence matrix. Sentence-term matrix is designed to represent diversity and redundancy within multiple articles. Finally, the top-weighted sentence in every cluster is picked out to form the summary until a user-preferred summary length is met. An evolutionary algorithm called Differential Evolution algorithm was also used to optimize data clustering process and could increase the quality of the generated text summaries (Abuobieda *et al.*, 2013b).

Some researchers employ clustering-based hybrid strategy to combine local and global search for sentence selection (Nie *et al.*, 2006). This approach does not depend only on similarity to cluster for sentence selection but also considers the overall document content similarity. In another related work, focus has been given on strengthening the clusters diversity. To achieve this, Aliguliyev (2010) used PSO algorithm by adding a mutation operation adopted from genetic algorithms to optimize intra-cluster similarity and inter-cluster dissimilarity.

Cluster based methods have been successful in its task to represent diversity and reduce redundancy within multiple articles. Although these can be considered the advantage of using clustering methods, as far as multi document is concerned, a summary cannot be meaningful enough if the relevance of a sentence is judged merely based on the clusters. This is because in clustering based method, eventually sentences are ranked according to its similarity with cluster centroid which simply represents frequent occurring terms.
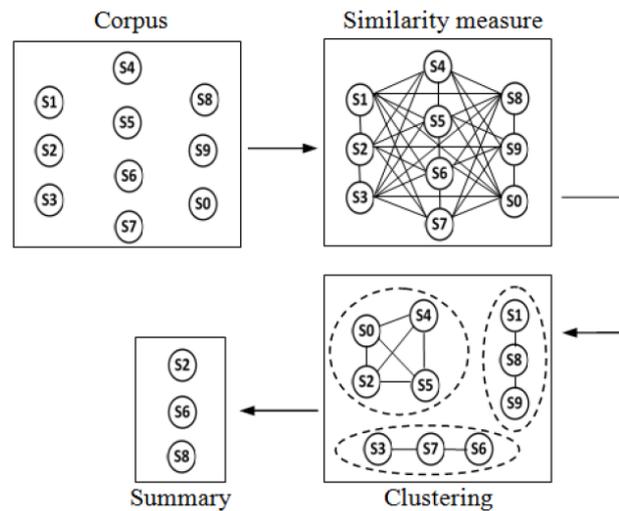

Fig. 4. A generalized architecture for cluster based summarization

Graph Based Method

Graph theory is simply used to model the connections or links that exist between objects. Generally, a graph can be denoted in the form of $G = (V, E)$, where V represents the graph's vertex or node and E is the edge between each vertex. In the context of text documents, vertex represents sentences and an edge is the weight between two sentences. Using this approach, documents can therefore be represented as a graph where each sentence becomes the vertex and the weight between each vertex corresponds to the similarity between the two sentences.

As in most literature concerning graph based approach, the most widely used similarity measure is the cosine similarity measure (Erkan and Radev, 2004). An edge then exists if the similarity weight is above some predefined threshold. Figure 5 shows an example graph for multi document. Once the graph is constructed for a set of documents, important sentences will then be identified; it follows the idea that a sentence is considered important if it is strongly connected to many other sentences.

This approach differs from the cluster based approach where sentences are ranked based on its closeness to cluster centroid. Two well-known graph based ranking algorithms is the HITS algorithm (Kleinberg, 1999) and the Google's PageRank (Brin and Page, 2012). Both methods have been traditionally used in Web-link analysis and social networks. Lexrank (Erkan and Radev, 2004) and TextRank (Mihalcea and Tarau, 2004) are two successful graph-based ranking systems that implement these algorithms.

Further studies have been carried to make improvement through modification in the ranking algorithm. Wan and Yang (2006) assigned different weights to intra-document links and inter-document links.

They gave more priority to sentence with high inter-document links. In later work by Hariharan and Srinivasan (2009), they approached the graph based method differently i.e., by discounting the already selected sentence by removing it from further consideration when they rank the remaining sentences in the document.

Apart from sentence level information, Wan (2008) and Wei *et al*. (2010) devised a document-sensitive graph model to explore document impact on the graph-based summarization, by incorporating both the document-level information and the sentence-to-document relationship in the graph-based ranking process. The document-level relations are used to adjust the weights of the vertices and the strength of the edges in the graph.

The approach to graph based methods have resulted in positive feedback from the multi document summarization research communities as it was able to identify 'prestigious' sentences across the documents. The resulting graph is also able to capture distinct topics from unconnected sub-graphs. However, this approach depends heavily on sentence similarity to generate graph, without "understanding" the relationship between the sentences.

Discourse Based Method

In this study, we also investigate studies related to discourse analysis. It involves analysis on the semantic relation that exist between textual units. In the case involving multiple document, some research works study the utility of cross-document relations to determine important sentences which are deemed relevant to the document collection.

Radev (2000), initiated the study on cross-document relations and came up with Cross-Document Structure Theory (CST) model. In this model, words, phrases or sentences can be link with each other if they are semantically connected. For example, some of the semantic connections or CST relations between sentences are given in Table 1.

Past studies have claimed that CST was indeed useful for document summarization. Zhang *et al*. (2002) have utilized CST to determine sentence relevance. First, they

produce multi document summary using a summarization system called MEAD (Radev *et al*., 2001). Then, they ask human experts to identify the CST relations that exist between sentences in the document set. At this point, the low scoring sentences are replaced with sentences that contains high CST relations. It was to produce summaries which are coherent; through the existence of relations between the summary sentences.

The effect of incorporating CST on the summarization process have likewise been contemplated by Jorge and Pardo (2010). They mainly investigate content selection methods for producing both informative and preference-based summaries. They tested their method using news articles acquired from CST News corpus (Aleixo and Pardo, 2008) which were annotated beforehand by human experts. The CST relations were utilized to treat repetition, complementarity and inconsistency among the diverse data sources. Nonetheless, the significant limitation of the above works is that the CST relations should be explicitly determined by human.
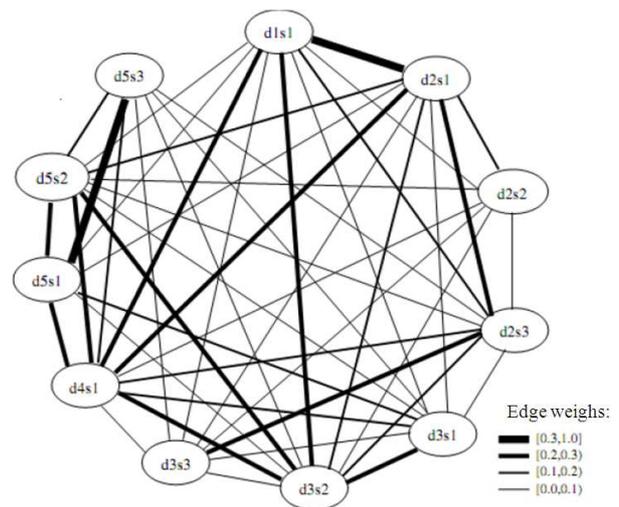

Fig. 5. Example graph as depicted in (Erkan and Radev, 2004). Each node represents a sentence

Table 1. Examples of CST relations (Zhang *et al*., 2002)

Relationship	Description	Text span 1 (S1)	Text span 2 (S2)
Identity	The same text appears in more than one location	Tony Blair was elected for a second term today.	Tony Blair was elected for a second term today.
Equivalence	Two text spans have the same information content	Derek Bell is experiencing resurgence in his career.	Derek Bell is having a comeback year.
Translation	Same information content in different languages	Shouts of "Viva la revolucion!" echoed through the night.	The rebels could be heard shouting, "Long live the revolution".
Subsumption	S1 contains all information in S2, plus additional information not in S2	With 3 wins this year, Green Bay has the best record in the NFL.	Green Bay has 3 wins this year.
Contradiction	Conflicting information	There were 122 people on the downed plane.	126 people were aboard the plane.
Historical background	S1 gives historical context to information in S2	This was the fourth time a member of the Royal Family has gotten divorced.	The Duke of Windsor was divorced from the Duchess of Windsor yesterday.

To address this gap, recent studies have attempted to identify the CST relations directly from texts document to produce summaries. Zahri and Fukumoto (2011) determined the CST relations by applying SVM classifier. The PageRank algorithm was used for sentence weighting whereby the directionality in PageRank was determined using the identified CST relations. Based on these relations, they also adjust the connected sentences to handle repetition issue.

In a similar study, Kumar *et al.* (2013) proposed Genetic-CBR classifier to identify CST relations from un-annotated documents. Two techniques based on voting model and fuzzy reasoning were used to rank the sentences (Kumar *et al.*, 2014). These techniques use the identified CST relationship between the sentences for sentence scoring. Both studies showed that CST based approach outperformed the cluster based method and graph based method.

Conclusion

In this study, the fundamental concepts and methods related to automatic text summarization have been discussed. Indeed, this study has been presented in a way that researchers new to this field are exposed to various automatic text summarization approaches and applications. The paper starts with a brief introduction to automatic text summarization and provides the review on past and present works found in the literature. Much discussion revolves around extractive based text summarization and primarily reviews approaches concerning sentence extraction, domain specific summarization and multi document summarization. It appears that each of the approaches discussed in this study possess its own advantages towards automatic summarization. However, there are a number of limitations pertaining to some approaches. Recent studies have attempted to address some of these limitations. The next big challenge is not only to focus on the summary information content, but efforts should also be put into the readability aspect of the generated summary itself. The future trend of automatic text summarization is most likely to move along this direction.

Acknowledgement

Author's Contributions

All authors equally contributed in this work.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Abuobieda, A., N. Salim, M.S. Binwahlan and A.H. Osman, 2013a. Differential evolution cluster-based text summarization methods. Proceedings of the International Conference on Computing, Electrical and Electronics Engineering, Aug. 26-28, IEEE Xplore Press, Khartoum, pp: 244-248. DOI: 10.1109/ICCEEE.2013.6633941

Abuobieda, A., N. Salim, Y.J. Kumar and A.H. Osman, 2013b. Opposition differential evolution based method for text summarization. In: Intelligent Information and Database Systems, Selamat, A., N.T. Nguyen, H. Haron (Eds.), Springer, Berlin, ISBN-10: 3642365434, pp: 487-496.

Aleixo, P. and T.A.S. Pardo, 2008. CSTNews: Um Córpus de Textos Journalísticos Anotados segundo a Teoria Discursiva CST (Cross-Document Structure Theory). Federal University of São Carlos.

Aliguliyev, R.M., 2009. A new sentence similarity measure and sentence based extractive technique for automatic text summarization. Expert Syst. Applic., 36: 7764-7772. DOI: 10.1016/j.eswa.2008.11.022

Aliguliyev, R.M., 2010. Clustering techniques and discrete particle swarm optimization algorithm for multi-document summarization. Comput. Intell., 26: 420-448. DOI: 10.1111/j.1467-8640.2010.00365.x

Aone, C., M.E. Okurowski, J. Gorlinsky and B. Larsen, 1999. A Trainable Summarizer with Knowledge Acquired from Robust NLP Techniques. In: Advances in Automatic Text Summarization, Mani, I. and M.T. Maybury (Eds.), MIT Press, Cambridge, ISBN-10: 0262133598, pp: 71-80.

Babar, S.A. and P.D. Patil, 2015. Improving performance of text summarization. Procedia Comput. Sci., 46: 354-363. DOI: 10.1016/j.procs.2015.02.031

Barzilay, R., K.R. McKeown and M. Elhadad, 1999. Information fusion in the context of multi-document summarization. Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, (LCL' 99), Stroudsburg, PA, USA, pp: 550-557. DOI: 10.3115/1034678.1034760

Becher, M., B. Endres-Niggemeyer and G. Fichtner, 2002. Scenario forms for web information seeking and summarizing in bone marrow transplantation. Proceedings of the Conference on Multilingual Summarization and Question Answering, (SQA' 02), Stroudsburg, PA, USA, pp: 1-8. DOI: 10.3115/1118845.1118846

Binwahlan, M.S., N. Salim and L. Suanmali, 2009. Swarm based text summarization. Proceedings of the International Association of Computer Science and Information Technology-Spring Conference, Apr. 17-20, IEEE Xplore Press, Singapore, pp: 145-150. DOI: 10.1109/IACSIT-SC.2009.61

Bossard, A. and C. Rodrigues, 2011. Combining a multi-document update summarization system–CBSEAS–with a genetic algorithm. Combinat. Intell. Methods Applic., 8: 71-87. DOI: 10.1007/978-3-642-19618-8_5

Brin, S. and L. Page, 2012. Reprint of: The anatomy of a large-scale hypertextual Web search engine. Comput. Netw., 56: 3825-3833. DOI: 10.1016/j.comnet.2012.10.007

Burges, C., T. Shaked, E. Renshaw, A. Lazier and M. Deeds *et al.*, 2005. Learning to rank using gradient descent. Proceedings of the 22nd International Conference on Machine Learning, (CML' 05), ACM, pp: 89-96. DOI: 10.1145/1102351.1102363

Conrad, J.G., J.L. Leidner, F. Schilder and R. Kondadadi, 2009. Query-based opinion summarization for legal blog entries. Proceedings of the 12th International Conference on Artificial Intelligence and Law, (AIL' 09), ACM, pp: 167-176. DOI: 10.1145/1568234.1568253

Daniel, N., D. Radev and T. Allison, 2003. Sub-event based multi-document summarization. Proceedings of the HLT-NAACL on Text Summarization Workshop, (TSW' 03), ACM, pp: 9-16. DOI: 10.3115/1119467.1119469

Edmundson, H.P., 1969. New methods in automatic extracting. J. ACM, 16: 264-285. DOI: 10.1145/321510.321519

Elhadad, N., M.Y. Kan, J.L. Klavans and K.R. McKeown, 2005. Customization in a unified framework for summarizing medical literature. Artifi. Intell. Med., 33: 179-198. DOI: 10.1016/j.artmed.2004.07.018

Erkan, G. and D.R. Radev, 2004. LexRank: Graph-based lexical centrality as salience in text summarization. J. Artifi. Intell. Res., 22: 457-479. DOI: 10.1613/jair.1523

Filatova, E. and V. Hatzivassiloglou, 2004. A formal model for information selection in multi-sentence text extraction. Proceedings of the 20th International Conference on Computational Linguistics, (CCL' 04), ACM, pp: 397-403. DOI: 10.3115/1220355.1220412

Fiszman, M., D. Demner-Fushman, H. Kilicoglu and T.C. Rindflesch, 2009. Automatic summarization of MEDLINE citations for evidence-based medical treatment: A topic-oriented evaluation. J. Biomed. Inform., 42: 801-813. DOI: 10.1016/j.jbi.2008.10.002

Fung, P. and G. Ngai, 2006. One story, one flow: Hidden Markov story models for multilingual multidocument summarization. ACM Trans. Speech Lang. Process., 3: 1-16. DOI: 10.1145/1149290.1151099

Galley, M., 2006. A skip-chain conditional random field for ranking meeting utterances by importance. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, (NLP' 06), ACM, pp: 364-372.

Ganesan, K., C. Zhai and J. Han, 2010. Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions. Proceedings of the 23rd International Conference on Computational Linguistics, (CCL' 10), ACM, pp: 340-348.

Gupta, V. and G.S. Lehal, 2010. A survey of text summarization extractive techniques. J. Emerg. Technol. Web Intell., 2: 258-268. DOI: 10.4304/jetwi.2.3.258-268

Hannah, M.E. and S. Mukherjee, 2014. A classification-based summarisation model for summarising text documents. Int. J. Inform. Commun. Technol., 6: 292-308. DOI: 10.1504/IJICT.2014.063217

Haque, M., S. Pervin and Z. Begum, 2013. Literature review of automatic multiple documents text summarization. Int. J. Innovat. Applied Stud., 3: 121-129.

Hariharan, S. and R. Srinivasan, 2009. Studies on graph based approaches for single and multi document summarizations. Int. J. Comput. Theory Eng., 1: 1793-8201.

Hariharan, S., 2010. Multi document summarization by combinational approach. Int. J. Comput. Cognit., 8: 68-74.

Hatzivassiloglou, V., J.L. Klavans and E. Eskin, 1999. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. Proceedings of the Joint Sigdat Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, (VLC' 99), pp: 203-212.

Hatzivassiloglou, V., J.L. Klavans, M.L. Holcombe, R. Barzilay and M.Y. Kan *et al.*, 2001. Simfinder: A flexible clustering tool for summarization. Columbia University Academic Commons.

Hovy, E. and C.Y. Lin, 1998. Automated text summarization and the SUMMARIST system. Proceedings of the Workshop on Held at Baltimore, (WHB' 98), Maryland, pp: 197-214. DOI: 10.3115/1119089.1119121

Hovy, E., 2005. Text Summarization. In: The Oxford Handbook of Computational Linguistics, Mitkov, R. (Ed.), OUP Oxford, Oxford, ISBN-10: 019927634X, pp: 583-598.

Hu, M., A. Sun and E.P. Lim, 2007. Comments-oriented blog summarization by sentence extraction. Proceedings of the 16th ACM Conference on Information and Knowledge Management, Nov. 06-10, Lisbon, Portugal, pp: 901-904. DOI: 10.1145/1321440.1321571

Jones, K.S., 1988. A Statistical Interpretation of Term Specificity and its Application in Retrieval. In: Document Retrieval Systems, Peter, W. (Ed.), Taylor Graham, London, pp: 132-142.

Jones, K.S., 1999. Automatic Summarising: Factors and Directions. In: Advances in Automatic Text Summarization, Mani, I., Maybury, M.T. (Eds.), MIT Press, Cambridge, ISBN-10: 0262133598, pp: 1-12.

Jorge, M.L.D.R.C. and T.A.S. Pardo, 2010. Experiments with CST-based multidocument summarization. Proceedings of the Workshop on Graph-Based Methods for Natural Language Processing, (NLP' 10), Stroudsburg, PA, USA, pp: 74-82.

Jurafsky, D. and J.H. Martin, 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. 1st Edn., Prentice Hall, Upper Saddle River, N.J., ISBN-10: 0131873210, pp: 988.

Kaicker, J., V.B. Debono, W. Dang, N. Buckley and L. Thabane, 2010. Assessment of the quality and variability of health information on chronic pain websites using the DISCERN instrument. BMC Med., 8: 59-59. DOI: 10.1186/1741-7015-8-59

Kaikhah, K., 2004. Text summarization using neural networks. Faculty Publications-Computer Science.

Kan, M.Y., K.R. McKeown and J.L. Klavans, 2001. Applying natural language generation to indicative summarization. Proceedings of the 8th European Workshop on Natural Language Generation, (NLC' 01), Stroudsburg, PA, USA, pp: 1-9. DOI: 10.3115/1117840.1117853

Khan, A., N. Salim, W. Reafee, A. Sukprasert and Y.J. Kumar, 2015. A clustered semantic graph approach for multi-document abstractive summarization. J. Teknol., 77: 61-72. DOI: 10.11113/jt.v77.6491

Khelif, K., R. Dieng-Kuntz and P. Barbry, 2007. An ontology-based approach to support text mining and information retrieval in the biological domain. J. Univ. Comput. Sci., 13: 1881-1907. DOI: 10.3217/jucs-013-12-1881

Klassen, P.P., 2012. Calculating LLR topic signatures with dependency relations for automatic text summarization. University of Washington.

Kleinberg, J.M., 1999. Authoritative sources in a hyperlinked environment. J. ACM, 46: 604-632.

Ko, Y. and J. Seo, 2008. An effective sentence-extraction technique using contextual information and statistical approaches for text summarization. Patt. Recognit. Lett., 29: 1366-1371. DOI: 10.1016/j.patrec.2008.02.008

Kogilavani, A. and B.D.P. Balasubramanie, 2009. Ontology enhanced clustering based summarization of medical documents. Int. J. Recent Trends Eng., 1: 546-549.

Kumar, Y.J., N. Salim, A. Abuobieda and A.T. Albaham, 2014. Multi document summarization based on news components using fuzzy cross-document relations. Applied Soft Comput., 21: 265-279. DOI: 10.1016/j.asoc.2014.03.041

Kumar, Y.J., N. Salim, A. Abuobieda and A. Tawfik, 2013. Multi document summarization based on cross-document relation using voting technique. Proceedings of the International Conference on Computing, Electrical and Electronics Engineering, Aug. 26-28, IEEE Xplore Press, Khartoum, pp: 609-614. DOI: 10.1109/ICCEEE.2013.6634009

Kupiec, J., J. Pedersen and F. Chen, 1995. A trainable document summarizer. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Jul. 09-13, Seattle, WA, USA, pp: 68-73. DOI: 10.1145/215206.215333

Lee, C.S., Z.W. Jian and L.K. Huang, 2005. A fuzzy ontology and its application to news summarization. IEEE Trans. Syst., Man Cybernet. Part B: Cybernet., 35: 859-880. DOI: 10.1109/TSMCB.2005.845032

Li, L., D. Wang, C. Shen and T. Li, 2010. Ontology-enriched multi-document summarization in disaster management. Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Jul. 19-23, Geneva, Switzerland, pp: 819-820. DOI: 10.1145/1835449.1835632

Luhn, H.P., 1958. The automatic creation of literature abstracts. IBM J. Res. Dev., 2: 159-165. DOI: 10.1147/rd.22.0159

Marcu, D., 1999. The automatic construction of large-scale corpora for summarization research. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 15-19, Berkeley, CA, USA, pp: 137-144. DOI: 10.1145/312624.312668

McKeown, K. and D.R. Radev, 1995. Generating summaries of multiple news articles. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Jul. 09-13, Seattle, WA, USA, pp: 74-82. DOI: 10.1145/215206.215334

McKeown, K.R., R. Barzilay, D. Evans, V. Hatzivassiloglou and J.L. Klavans *et al.*, 2002. Tracking and summarizing news on a daily basis with Columbia's Newsblaster. Proceedings of the 2nd International Conference on Human Language Technology Research, (LTR' 02), ACM, pp: 280-285.

McKeown, K., J. Klavans, V. Hatzivassiloglou, R. Barzilay and E. Eskin, 1999. Towards multidocument summarization by reformulation: Progress and prospects. Proceedings of the 16th National Conference on Artificial Intelligence and 11th Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence, (AAI' 99), ACM, pp: 453-460.

Mihalcea, R. and P. Tarau, 2004. TextRank: Bringing order into texts. Proceedings of the Conference on Empirical Methods in Natural Language Processing, (NLP' 04), University of Michigan, pp: 1-8.

Naderi, N. and R. Witte, 2010. Ontology-based extraction and summarization of protein mutation impact information. Proceedings of the Workshop on Biomedical Natural Language Processing, (NLP' 10), ACM, pp: 128-129.

Nenkova, A. and A. Bagga, 2004. Facilitating email thread access by extractive summary generation. Hohn Benjamins Publishing Company. DOI: 10.1075/cilt.260.32nen

Nenkova, A. and K. McKeown, 2011. Automatic summarization. Foundat. Trends Inform. Retrieval, 5: 103-233.

Nenkova, A. and K. McKeown, 2012. A Survey of Text Summarization Techniques. In: Mining Text Data, Aggarwal, C.C. and C. Zhai (Eds.), Springer Science and Business Media, New York, ISBN-10: 1461432235, pp: 43-76.

Nenkova, A. and L. Vanderwende, 2005. The impact of frequency on summarization, Microsoft Research.

Nenkova, A., L. Vanderwende and K. McKeown, 2006. A compositional context sensitive multi-document summarizer: Exploring the factors that influence summarization. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 06-10, Seattle, WA, USA, pp: 573-580. DOI: 10.1145/1148170.1148269

Neto, J.L., A.A. Freitas and C.A.A. Kaestner, 2002. Automatic Text Summarization using a Machine Learning Approach. In: Advances in Artificial Intelligence, Bittencourt, G. and G.L. Ramalho, Springer, pp: 205-215. DOI: 10.1007/3-540-36127-8_20

Newman, P.S. and J.C. Blitzer, 2003. Summarizing archived discussions: A beginning. Proceedings of the 8th International Conference on Intelligent user Interfaces, Jan. 12-15, Miami, FL, USA, pp: 273-276. DOI: 10.1145/604045.604097

Nie, Y., D. Ji, L. Yang, Z. Niu and T. He, 2006. Multi-document summarization using a clustering-based hybrid strategy. Inform. Retrieval Technol., 4182: 608-614. DOI: 10.1007/11880592_53

Radev, D.R. and K.R. McKeown, 1998. Generating natural language summaries from multiple on-line sources. Comput. Linguist., 24: 470-500.

Radev, D.R., 2000. A common theory of information fusion from multiple text sources step one: Cross-document structure. Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue, (WDD' 00), ACM, pp: 74-83. DOI: 10.3115/1117736.1117745

Radev, D.R., H. Jing, M. Styś and D. Tam, 2004. Centroid-based summarization of multiple documents. Inform. Process. Manage., 40: 919-938. DOI: 10.1016/j.ipm.2003.10.006

Radev, D.R., S. Blair-Goldensohn and Z. Zhang, 2001. Experiments in single and multi-document summarization using MEAD. Ann. Arbor, 1001: 48109-48109.

Rambow, O., L. Shrestha, J. Chen and C. Lauridsen, 2004. Summarizing email threads. Proceedings of HLT-NAACL: Short Papers, (HSP' 04), ACM, pp: 105-108.

Saggion, H. and T. Poibeau, 2013. Automatic Text Summarization: Past, Present and Future. In: Multi-Source, Multilingual Information Extraction and Summarization, Poibeau, T., H. Saggion, J. Piskorski and R. Yangarber (Eds.), Springer Science and Business Media, Berlin, ISBN-10: 3642285694, pp: 3-21.

Schilder, F. and R. Kondadadi, 2008. FastSum: Fast and accurate query-based multi-document summarization. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, (TSP' 08), pp: 205-208.

Schilder, F., R. Kondadadi, J.L. Leidner and J.G. Conrad, 2008a. Thomson reuters at tac 2008: Aggressive filtering with fastsum for update and opinion summarization. Proceedings of the 1st Text Analysis Conference, (TAC' 08).

Schilder, F., J.L. Leidner, J.G. Conrad and R. Kondadadi, 2008b. Polarity filtering for sentiment summarization. Proceedings of the 1st Text Analysis Conference, (TAC' 08) NIST, Gaithersburg, MD.

Suanmali, L., N. Salim and M.S. Binwahlan, 2009. Fuzzy logic based method for improving text summarization. Int. J. Comput. Sci. Inform. Security, 2: 1-6.

Suanmali, L., N. Salim and M.S. Binwahlan, 2011. Fuzzy genetic semantic based text summarization. Proceedings of the IEEE 9th International Conference on Dependable, Autonomic and Secure Computing, Dec. 12-14, IEEE Xplore Press, Sydney, NSW, pp: 1184-1191. DOI: 10.1109/DASC.2011.192

Sundheim, B.M., 1992. Overview of the fourth message understanding evaluation and conference. Proceedings of the 4th Conference on Message Understanding, (CMU' 92), ACM, pp: 3-21. DOI: 10.3115/1072064.1072066

Svore, K.M., L. Vanderwende and C.J. Burges, 2007. Enhancing single-document summarization by combining Ranknet and third-party sources. Microsoft Corporation.

Verma, R., P. Chen and W. Lu, 2007. A semantic free-text summarization system using ontology knowledge.

Wan, X. and J. Yang, 2006. Improved affinity graph based multi-document summarization. Proceedings of the Human Language Technology Conference of the NAACL, ACM, pp: 181-184.

Wan, X., 2008. An exploration of document impact on graph-based multi-document summarization. Proceedings of the Conference on Empirical Methods in Natural Language Processing, (NLP' 08), ACM, pp: 755-762.

Wei, F., W. Li, Q. Lu and Y. He, 2010. A document-sensitive graph model for multi-document summarization. Know. Inform. Syst., 22: 245-259. DOI: 10.1007/s10115-009-0194-2

White, M., T. Korelsky, C. Cardie, V. Ng and D. Pierce *et al*., 2001. Multidocument summarization via information extraction. Proceedings of the 1st International Conference on Human Language Technology Research, (LTR' 01), pp: 1-7. DOI: 10.3115/1072133.1072206

Wu, C.W. and C.L. Liu, 2003. Ontology-based text summarization for business news articles. Proceedings of the ISCA 18th International Conference on Computers and Their Applications, (CTA' 03), pp: 389-392.

Xia, Y., Y. Zhang and J. Yao, 2011. Co-clustering Sentences and Terms for Multi-Document Summarization. In: Computational Linguistics and Intelligent Text Processing, Springer Science and Business Media, Berlin, ISBN-10: 3642193994, pp: 339-352.

Zahri, N.A.H.B. and F. Fukumoto, 2011. Multi-document summarization using link analysis based on rhetorical relations between sentences. In: Computational Linguistics and Intelligent Text Processing, Springer Science and Business Media, Berlin, ISBN-10: 3642193994, pp: 328-338.

Zhang, Z., S. Blair-Goldensohn and D.R. Radev, 2002. Towards CST-enhanced summarization. Proceedings of the 18th National Conference on Artificial Intelligence, (CAI' 02), pp: 439-446.

Zhou, L. and E.H. Hovy, 2006. On the summarization of dynamically introduced information: Online discussions and blogs. American Association for Artificial Intelligence.