International Journal of English Linguistics; Vol. 7, No. 4; 2017
ISSN 1923-869X E-ISSN 1923-8703
Published by Canadian Center of Science and Education

Addressing the Problem of Coherence in Automatic Text Summarization: A Latent Semantic Analysis Approach

Abdulfattah Omar[1&2]

[1] Department of English, College of Science and Humanities, Prince Sattam Bin Abdulaziz University, Al Kharj, Saudi Arabia

[2] Department of English, Faculty of Arts, Port Said University, Urban, Egypt

Correspondence: Abdulfattah Omar, Department of English, College of Science and Humanities, Prince Sattam Bin Abdulaziz University, Al Kharj, Saudi Arabia. E-mail: a.abdelfattah@psau.edu.sa

Received: March 13, 2017 Accepted: April 1, 2017 Online Published: July 15, 2017

doi:10.5539/ijel.v7n4p33 URL: http://doi.org/10.5539/ijel.v7n4p33

Abstract

This article is concerned with addressing the problem of coherence in the automatic summarization of prose fiction texts. Despite the increasing advances within the summarization theory, applications and industry, many problems are still unresolved in relations to the applications of the summarization theory to literature. This can be in part attributed to the peculiar nature of literary texts where standard or typical summarization processes are not amenable for literature. This study, therefore, tends to bridge the gap between literature and summarization theory by proposing a summarization system that is based on more semantic-based approaches for extracting more meaningful and coherent summaries. Given that lack of coherence within summaries has its negative implications on understanding original texts; it follows that more effective methods should be developed in relation to the extraction of coherent summaries. In order to do this, a hybrid of methods including statistical (TF-IDF) and semantic (Latent Semantic Analysis LSA) methods were used to derive the most distinctive features and extract summaries from 10 English novellas. For evaluation purposes, both intrinsic and extrinsic methods are used for determining the quality of the extracted summaries. Results indicate that the integration of LSA into features extraction methods achieves better summarization performance outcomes in terms of coherence properties within the extracted summaries.

Keywords: automatic summarization- cohesion- coherence- extraction- Latent Semantic Analysis- TF-IDF

1. Introduction

The recent explosive growth of digital and online texts has posed a number of challenges in text summarization research. Traditionally this process has been paper-based using what can be described as the philological method where researchers and professionals tended to read source texts and compose their own summaries based on the selection of what they think to be the most significant sentences within these texts. The advent of electronic text, however, has raised many issues concerning the reliability and effectiveness of these traditional methods. The prolific size of digital corpora as well as the complexity of data abstracted from them make it imperative today to develop more reliable methods that can deal with these challenges in an effective way. Recognizing the ineffectiveness of manual and traditional methods, researchers are increasingly turning to computational and machine-based methods for carrying out summarization tasks. Over the recent years, different methods have been proposed in the study of automatic text summarization (ATS) including extraction-based, abstraction-based, and aided summarization methods. However, extraction methods remain the most widely used so far. These have largely been based on generating summaries in the form of generic extracts; that is, the resulting document summary is a sequence of fragments of the original text. Generally speaking, these methods depend on statistical/ quantitative weighting methods for extracting the most distinctive words and phrases. One problem with think kind of summarization, however, is that sentence relevance is not always accurate. Summaries of this kind suffer a very serious problem which is lack of sentence relevance. Therefore, summaries extracted are not coherent. This problem is even more challenging in the summarization of literary texts including novels and short fiction where extracted summaries cannot express well what texts are about. In most cases, the summaries do not reveal the development of actions and do not give the reader the expected information about a given text. The implication is that more coherent summaries are required for finding the main ideas of a text as well as

capturing the author's concepts, yet in a more cohesive and coherent manner. Part of the problem is related to the peculiar nature of literature in natural language processing applications. It is also true that the applications of digital and computer technology to literature are still very limited. In the face of this problem, I argue that more semantic-based approaches are required for generating more reliable summaries. The study proposes the integration of Latent Semantic Analysis (LSA) methods into text summarization where cohesive properties are detected and exploited for building more relevant sentences and identifying local coherence structures throughout the whole text with the purpose of generating more coherent and meaningful summaries. LSA is based on identifying the semantically important sentences using semantic normalization of topic space and further weighting of each topic using sentences representation in topic space. In order to examine the effectiveness of the proposed method, 10 English prose fiction texts were selected for the purpose.

The remainder of this article is organized as follows. Part 2 outlines the problem of the research. Part 3 asks the research questions. Part 4 is a brief survey of summarization literature. Part 5 is research methodology. It defines methods, data and procedures. Part 6 reports the results. Part 7 is conclusion.

2. Statement of the Problem

ATS is a process where a computer summarizes a text using a software program usually referred to as a summarization system (Hovy, 2005; Mani, 2001; Mani & Maybury, 1999; Tan, 2012; Torres-Moreno, 2014). In these systems, a text is given to the computer and the computer generates a shorter version of the original text (Mani, 2001; Patil et al., 2015). The main function of any summarization system is to help the user to find the needed information and to present the content of the source document in as compact a style as possible, i.e. as a summary. In this way, it is a reduction process in the first place. Saggion & Piobeau (2012, p. 3) argue that automatic summarization is essentially a "computer-based production of condensed versions of documents".

Text summaries are very important for familiarizing oneself with a subject matter and saving time. Summaries have become more important now since the availability of and ease accessibility to information have impressed our daily life. The flood of online digital information and the growth of the World Wide Web have made the notion of information more important in modern societies. In academic contexts, for researchers, it is very difficult for researchers to real all the materials on a given subject. ATS can help researchers get the key points of a certain topic from a large amount of literature in an efficient way. Text summaries are helpful in this way since they can help users digest information content. Then he can easily determine the more relevant documents without reading the whole documents. In other words, text summarization is helpful for dealing with this information overload by automatically generating summaries that give the gist of original documents (Badry et al., 2013). It is not surprising then that automatic research and applications in the summarization of scientific articles receives a high priority of researchers and professionals. In non-academic contexts, ATS is used in different and numerous domains either in professional contexts or daily life. ATS is an active area in business and legal environments where summarizations systems are widely used in producing summaries of meeting minutes and legal documents. Summarizers are even used for generating summaries of web pages and email threads. Therefore, summarizers help millions of people to be updated without having to read all the materials. Without summarization, it would be impossible for people today to be updated with that growing mass of information accessible online.

Despite the development of many summarizers, some fundamental problems remain unsolved. Lack of coherence is a recurring problem in the automatic summarization performance due to the fact that there is no sentence relevance. The sentences or clauses in summaries are not usually connected to each other and do not support the overall argumentative structure of the text. In this, the thematic significance of original texts is usually not considered in the extracted summaries which follows that extracted summaries are sometimes misleading for readers and users. This problem is attributed to the way these summaries are usually extracted. Generally, summaries of the kind are generated based on weighting methods where the most frequent words and phrases are kept. In this, the relation between phrases and sentences is not considered. The result is that summarizers come up with a summary that may have the most important phrases and sentences but are not well connected to each other. The implication of this problem on automatic classification is that lack of coherence within the automatic summaries has its negative effects on understanding the original text. Incoherent summaries are misleading for readers if they do not have access to original texts (Foltz et al., 1998; Lapata & Barzilay, 2005; Porzel, 2010; Rico-Jimenez, 2016).

The claim here is that coherence and sentence relevance remain unsolved challenges for almost all text summarization systems. This can be attributed to the fact that much of the summarization literature and industry are still using surface information methods for deriving meanings out of texts with no deep semantic analysis. By

coherence here, I mean, the relation between sentences and clauses and how these form a coherent and meaningful structure. The problem is more challenging in the summarization applications on literary texts including novels and short fiction. The lack of sentence relevance and coherence properties has negative implications on the adequate representation of texts for summarization tasks and thus summary readability. It is also argued that literary texts need to be addressed differently. The conventional or typical methods that are used in relation to other sorts of data such as news articles, and legal documents cannot be appropriate in producing meaningful and coherent summaries in relation to literary texts including novels and short stories. In the typical summarization systems, for instance, the title of the document and location of sentences and phrases within the document are considered indicators of salient features within documents. These, however, do not necessarily indicate any significance in novels and short stories. In this context, this study suggests more semantic methods for the extraction of distinctive phrases and sentences. The rationale is that sentence relevance is a meaning issue which needs to be addressed using semantic approaches. To put it into effect, the study proposes the integration of latent semantic analysis (LSA) methods into automatic summarization for generating more reliable summarization performance (Landauer, 2007).

3. Research Questions

In the light of the above mention problem, this article asks the following research questions: (1) do latent semantic analysis methods achieve useful summarization performance in relation to coherence and sentence relevance? And (2) are the resulted summaries based on coherent sentences and clauses which reveal the thematic significance of the literary texts? In order to answer the research questions, LSA methods are suggested for improving the quality of summarization performance in relation to literary texts. The study is based on a corpus of 10 English novellas and comparing them to their manual summaries. The objective is to produce summaries that are meaningful and coherent. By coherent I mean summaries that are based on relevant sentences and which are not misleading for readers and users.

4. Previous Work

ATS has been developed over the last five decades with the purpose of generating automatic summaries via the computer. The initial developments of the approach were in the 1950s. Perhaps the most cited paper on summarization is that of (Luhn, 1958). However, the recent decades have witnessed a great and unprecedented development in text summarizers (Sylva, 2015; Wang et al., 2010; Yeh, 2005). An important development of the ATS came with Microsoft Word in 1997 where the corporation proposed its first summarizer for documents. Numerous summarizers have been introduced; many of them were featured with commercial aspects. These include StarOffice Summarizer, Copernic Summarizer, TextRank, Microsoft Office Word Summarizer, and OpenOffice Summarizer.

In text summarization, a system is supposed to keep only the most distinctive words, phrases, and sentences. The failure to do so has negative implications on the summarization performance and reliability. The effectiveness of a summarization system is thus depends on extracting only and all the important information within a text (Juan-Manuel & Torres-Moreno, 2014). This has been conventionally done using different weighting methods that can statistically weigh the distinctiveness of words, phrases, and sentences. These can be categorized under three main approaches, namely variance, term frequency analysis, and Principal Component Analysis (PCA) methods.

In variance analysis, high dimensionality of data can be effective in keeping the most varied words and phrases within a text. The assumption is that variables describing the characteristics of interest are thus only useful for summarization if there is significant variation in the values they take (Chua & Asur, 2013; Ferreira; 2013; Mani, 2001; Moreno, 2016). In spite of its effectiveness in keeping only the features that are significant in relation to variation, it is not the only one factor that needs to be taken into account. Therefore, different studies tend to use frequency analysis along with variance analysis. Frequency analysis methods are thus used for identifying the most important features (terms) within documents. The underlying principle if descriptions are longer, terms will be used more often. Therefore, they need to be kept in summarization processes. This assumption can be, however, falsified. Lee et al. (2003) argue that a term which occurs more frequently is not necessarily a good discriminator, and should be given less weight than one which occurs less frequently. In order to overcome this problem, PCA has been extensively used in summarization tasks whereby a summary is generated by extracting sentences that are likely to represent the main theme of a document (Bhatia & Jaiswal, 2015; Canhasi & Kononenko; 2016; Kogilavani, 2016).

This is one of the basic geometric tools that are used to produce a lower number of the vectors within a corpus (Härdle & Simar, 2003; Jackson, 1991). The main function of PCA is to find the most informative vectors within

a data matrix. Jolliffe (2002) explains "The central idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data sets (2002, p. 1). It can be thus described as a technique for data quality (Jackson, 1991). To put it simply, PCA performs two complementary tasks: (1) organizing sets of data and (2) reducing the number of variables without much loss of information. In many automatic summarization applications, PCA is used for reducing the number of variables (words and phrases) so that summarization is based on the most distinctive vectors or features within data sets. The literature suggests that PCA is used a great deal in automatic summarization applications prior to executing the summarization task for finding patterns in data that are built on uncorrelated vectors. In spite of the computational mathematical nature of PCA, this discussion is only concerned with the idea of data reduction and its implications to document summarization.

The main assumption behind PCA is that a text or corpus with huge data sets can be reduced so that the most distinctive vectors are identified with the purpose of best expressing the data and revealing hidden structures. Although some of the discarded or deleted variables can be important within the targeted texts, PCA works to perform a 'good' dimensionality reduction with no great loss of information. The underlying principle of PCA is that it removes correlated variables within datasets so that it describes the covariance relationships among these variables. In this way, PCA has proved effective in retaining the most distinctive features and also in the summarization of related documents, which is referred to in the literature as multi-document summarization (Fiori, 2014; Ketui, 2014; Li, 2015; Poibeau, 2013; Zhuge, 2016). Nevertheless, it is evident that extracted summaries still lack coherence. There is no relevance between the retained phrases and sentences.

The implication here is that the different weighting methods can be effective in retaining the most distinctive and important features within documents as well as discarding information of secondary importance which lead to generating summaries based on the most important information within documents. Nevertheless, these statistical methods are not effective in dealing with the problem of coherence within the resulted summaries which suggests that semantic-based methods should be integrated into summarization systems for improving sentence relevance and thus producing more coherent summaries.

In different summarization systems, weighting methods have always been combined with other methods that worked together to identify the most distinctive features within documents. One of these is the use of sentence location (Fattah, 2014; Nenkova & McKeown, 2011; Shah & Jivani, 2016). The assumption is that there is a close relation between the position of a sentence in a text and how much information it has. In this way, it is suggested that important sentences come at the beginning and end of documents where it is most likely to find topic and concluding sentences. Other methods included classifying and summarizing similar texts together in what is referred to as multi-document summarization. To illustrate the argument, let's take this example. Given a set of scientific articles on generative grammar, these documents are more likely to share some information which should be judged as important. Multi-document summarization processes then will lead to producing summaries that are based on the most important information within these documents.

Parallel to the development of different summarization systems and methods, different evaluation methods and approaches have been devised to evaluate the usefulness of such systems and methods. Some of these have been concerned with addressing the issue of sentence relevance and coherence in summaries (Fiori, 2014; Mashechkin et al., 2011; Nenkova & McKeown, 2011; Wang et al., 2010). Evaluation systems almost agree that existing summarization systems need to integrate semantic-based methods for improving the quality of summarization performance and building more coherent and meaningful summaries.

The literature however suggests that very little has been done in relation to the summarization of literary texts. Not surprisingly, much of the attention has been paid to the summarization of daily news and news articles due to the prolific size of news generated every day and people's need to be updated with news from here and there. According to (Nenkova & McKeown, 2011), scientific papers, medical and legal documents come next. In these genres, there is usually a well-established structured that makes it more appropriate to generate coherent summaries. In research papers, for instance, there is usually an introduction, literature review, methodology, discussion, and conclusion parts where authors are concerned with clarity and coherence. In documents of the kind, there is usually no ambiguity. Therefore, it is usually easy for text summarizers to produce summaries that are clear and coherent. In literature, on the other hand, we do not usually have that well-established structure. In other words, there is no template for writing prose fiction in spite of the fact that certain elements including character, plot, character, setting, dialogue, and point of view that need to be considered in writing novels, novellas, and short stories. Furthermore, these are usually based on metaphoric language which makes it infeasible to have summaries that are meaningful and coherent. In this way, automatic summarization processes of literary texts are really difficult and challenging. This may explain the idea that summaries of literary texts are

still done using traditional manual methods. In the digital age we live, this seems unreliable. It is very challenging to produce summaries for all literary works today in an efficient manner. Libraries, digital libraries, and archives, for instance, need more reliable ways for generating summaries or synopses of the literary texts so that readers know what texts are about. This is important since usually need exact information about published texts to determine whether they read them or not. The implication here is that since typical automatic summarization processes are not appropriate for literary texts, these need to be addressed in a different way due to the peculiar nature of literary texts. This article tends to address this gap in literature by finding ways that can build summaries of prose fiction texts that are both meaningful and coherent.

5. Methodology

5.1 Methods

The proposed method proposes the integration of latent semantic analysis methods into summarization processes for addressing the problem of coherence within the automatic summarization of prose fiction. Latent Semantic Analysis (LSA) is an approach that is concerned with analyzing documents with the purpose of identifying the underlying meanings or concepts within these documents. It was originally developed for extracting and representing the underlying semantic connections between both the documents and the words in a large corpus of texts for the purpose of automatic indexing or grouping of documents (Adrian et al., 2007; Deerwester et al., 1990; Dumais et al., 1988; Foltz et al., 1998; Landauer et al., 1998).

The literature suggests that LSA was originally developed to tackle some problems such as polysemy and synonymy that used to affect the validity of automatic classification performance. Today, it has numerous applications and techniques. Almost all LSA models assume in principle that a document arises from one single source even if that source is not determined or defined. The underlying principle of LSA is that it uses statistical correlation between word and passage meaning to create a similarity score between any two documents based entirely on the words that they contain. Landauer et al. (1998) assert that the relations LSI generates are well correlated with several human cognitive phenomena involving association or semantic similarity. LSA "uses as its initial data not just the summed contiguous pairwise (or tuple-wise) co-occurrences of words but the detailed patterns of occurrences of very many words over very large numbers of local meaning-bearing contexts, such as sentences or paragraphs, treated as unitary wholes" (Landauer et al., 1998, p. 5). The effectiveness of LSA in automatic grouping makes it possible to use it in the automatic summarization of literary texts.

The implication is that the use of LSA will be useful in improving cohesion and coherence properties within the summaries. In ATS, cohesion refers to the relationships among the elements of a text while coherence refers to relationships between text segments (Bhatia & Jaiswal, 2016; Cha & Kim, 2016; Gambhir & Gupta, 2017; Jaradat & Al-Taani, 2016). It should be also clear that the proposed system considers the issue of keeping the sentences and clauses that are central to the story. This is best described in terms of salience. This is a property within ATS performance where a summary is produced in a way that preserves only the important information, discards secondary information, and considers the relevance of the content (Mani, 2001). The objective, after all, is to generate summaries that provide the reader with a coherent idea of what the text (whether it is a novel or short story) is about.

In our case, LSA methods are used to identify the topical shifts within the documents which are thought to be indicators of salience and significance. According to Kazantseva & Szpakowicz (2012; 2014), topical shifts are characterized by changes in the vocabulary used by the author. The assumption therefore is that identifying these lexical properties will be useful in maintaining the cohesive properties within the summary.

5.2 Data

This study is based on a corpus of 10 English novellas written by British and American writers. A novella is a distinct narrative form of prose fiction that is normally longer than a short story and shorter than a novel (Gillespie, 1967; Kercheval, 1997; Leibowitz, 1974). The rationale of selecting the novella genre in particular is that they are appropriate for experimenting the proposed approach. The length of the novellas as well is appropriate for validating processes. The results of the study will then be applicable to either short stories or novels since the novella combines the features of both. The main criteria of selecting the novellas are that they are roughly of the same length and that they are conventional or typical novellas. By conventional I mean that they have the elements of novellas including characters, setting, and point of view. The study also avoided experimental texts as they have a different nature which entails that they may need to be addressed differently. The selected texts undergo a process of single-document summarization using LSA methods. Details of the corpus and procedures are described below. The texts are chronologically ordered as follows.

The selected texts include Charles Dickens' *A Christmas Carol* (1843), Thomas Hardy's *An indiscretion in the Life* of *an Heiress* (1878), Henry James' *Daisy Miller* (1879), H. G. Wells' *The War of the Worlds* (1898), Joseph Conrad's *Heart of Darkness* (1899), Albert Camus' *The Stranger* (1942), George Orwell's *Animal Farm* (1945), Ernest Hemingway's *The Old Man and the Sea* (1952), Richard Matheson's *I Am Legend* (1954), Nora Ephron *Heartburn* (1983).

5.3 Procedures

In order to generate summaries that are both meaningful and coherent of the selected texts, a summarization system is developed. The system is designed to work in two subsequent stages. As an initial step, TF-IDF is used for the identification of the most significant or distinctive thematic features of the selected documents.TF-IDF is now one of the most common methods for identifying the most important variables within datasets (Robertson, 2004). TF-IDF works on what is described as the specificity principle. According to Jones (1972, p. 11), this is "a semantic property of index terms: a term is more or less specific as its meaning is more or less detailed and precise". The underlying principle of specificity is the selection of particular terms, or rather the adoption of a certain set of effective vocabulary that collectively characterizes the set of documents. In TF-IDF, the most discriminant terms, phrases and sentences are the highest TF-IDF variables. This is computed by summing the TF-IDF for each query term and a high weight in TF-IDF is reached by a high term frequency in the given document and a low document frequency of the term in the whole collection of documents (Salton & Buckley, 1987; Salton & Buckley, 1988). The implication to automatic summarization is that if the highest TF-IDF variables, which are taken to be the most discriminant terms, are identified, then unimportant variables can be deleted and data dimensionality is reduced. In this way, the summary will be based on only distinctive and significant features.

After the selection of the highest TF-IDF variables, LSA methods will be integrated in order to make sure that retained sentences are relevant. This represents the second stage in the proposed summarization system. At this stage, topical shifts are identified with the purpose of having a meaningful background about each story. The system is also trained to identify the temporal expressions, discourse markers, and transition words that are used within the documents as cohesive devices. This technique works in two directions. First, it identifies the event sentences within the texts. Second, it identifies the cohesive elements within these texts. This will help in extracting a summary that best expresses the events and development of the action within the novellas. The identification and extraction of transitional words and discourse markers will be useful in connecting the parts of the extracted summary together. The summarization system is built using GATE (General Architecture for Text Engineering) software. This is open source software that is widely used for text processing applications. The software has been proved useful and effective in dealing with solving different text processing problems (Cunningham et al., 2002; Cunningham et al., 1999; Rutkauskas & Bargelis, 2016). Finally, both intrinsic and extrinsic measures are used for evaluating the usefulness and meaningfulness of the generated automatic summaries. Using intrinsic evaluation methods, all the extracted summaries are compared to summaries written by people who teach English novel. Three people were asked to extract the most important 50 sentences within each document. They were told to write summaries that best describe the development of actions and give information about the text. During this process, each participant was asked to produce an extracted summary for each of the selected 10 texts. For a novella type, it is thought that 50 sentences will be appropriate for providing the reader a clear image of what the book is about. For short stories, it can be shorter and for a novel it can be a bit longer. The purpose here anyway is to compare the automatic summaries to these human made summaries in order to see how similar they are using a similarity score. The general rule is that the higher similarity score is, the closer automatic summaries are to human ones. As a final step, judges were asked to evaluate the extracted summaries in terms of content, cohesion, and coherence. They were asked to say whether these summaries indicate the content of the texts in a meaningful and coherent way.

6. Results

In the process of extracting summaries of the selected texts, eventual clauses were identified and retained with the purpose of giving a background of each story so that the reader knows what it is about. The rationale is to select only the clauses that tell the important events within the texts. Discourse markers and transition expressions were also signalled and identified in order to build cohesive relations between the extracted clauses.

For evaluation purposes, the automatic summaries extracted here are compared to those produced by the participants using manual methods. As above mentioned, the process involved three professionals who extracted a summary of each of the selected texts. Each of the extracted automatic summaries was compared to the three manual summaries produced by the participants. The purpose of this comparison here is to determine whether

the automatic summaries are based on the most distinctive clauses.

As an initial step, the manual summaries produced by the three participants were revealed to have a high degree of sentence overlap. This is shown as follows.

Table 1. Sentence overlap in the manual summaries

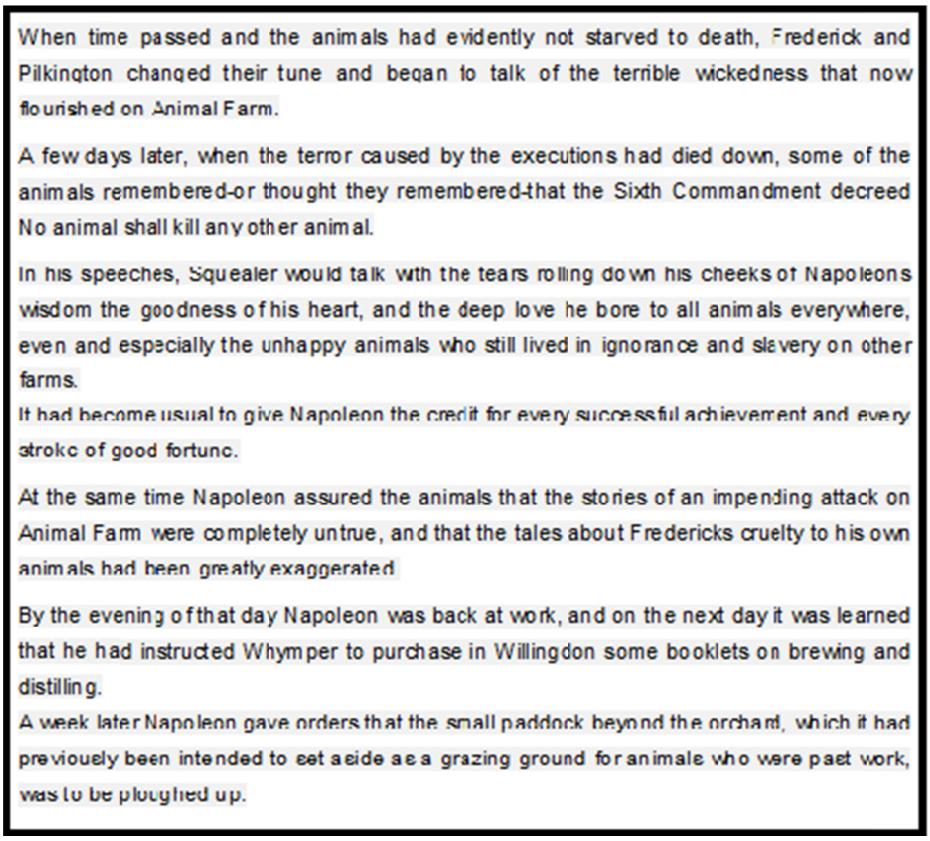
No.	Document title	Document code	Number of overlapping sentences and clauses
1.	*A Christmas Carol*	**001**	37
2.	*An indiscretion in the Life* of *an Heiress*	**002**	35
3.	*Daisy Miller*	**003**	29
4.	*The War of the Worlds*	**004**	31
5.	*Heart of Darkness*	**005**	38
6.	*The Stranger*	**006**	36
7.	*Animal Farm*	**007**	34
8.	*The Old Man and the Sea*	**008**	33
9.	*I Am Legend*	**009**	39
10.	*Heartburn*	**010**	41

As above shown, the similarity scores for the documents are close to each other. They range from 29 to 41. For convenience reasons, the three participants were asked to produce one joint summary for each of the selected texts in order to be compared to the automatic summaries. Results indicate that there is high similarity between the extracted automatic summaries and the joint manual summaries JMS. Agreement measures based on computing sentence overlap of the automatic summaries and the JMS can be summarized as follows.

Table 2. Sentence overlap in the JMS and automatic summaries

Document code	Document title	Number of overlapping sentences and clauses
001	*A Christmas Carol*	34
002	*An indiscretion in the Life* of *an Heiress*	30
003	*Daisy Miller*	30
004	*The War of the Worlds*	28
005	*Heart of Darkness*	33
006	*The Stranger*	32
007	*Animal Farm*	32
008	*The Old Man and the Sea*	28
009	*I Am Legend*	31
010	*Heartburn*	35

By the way, when the automatic summaries were compared to each of the participants' summaries separately, similarity score was higher. The claim here is that the proposed system based on TF-IDF and LSA methods is effective in retaining what thought to be the most distinctive features within each document. The extracted sentences are central to the development of the action. There is a high degree of agreement between the manual summaries on the one hand and the automatic ones on the other. Concerning coherence among the elements and clauses of the summary, results indicate that summaries are meaningful and coherent. The proposed system is effective in identifying the eventual clauses that best describe the development of the action. This can be illustrated in the figure below.


When time passed and the animals had evidently not starved to death, Frederick and Pilkington changed their tune and began to talk of the terrible wickedness that now flourished on Animal Farm.

A few days later, when the terror caused by the executions had died down, some of the animals remembered-or thought they remembered-that the Sixth Commandment decreed No animal shall kill any other animal.

In his speeches, Squealer would talk with the tears rolling down his cheeks of Napoleons wisdom the goodness of his heart, and the deep love he bore to all animals everywhere, even and especially the unhappy animals who still lived in ignorance and slavery on other farms.

It had become usual to give Napoleon the credit for every successful achievement and every stroke of good fortune.

At the same time Napoleon assured the animals that the stories of an impending attack on Animal Farm were completely untrue, and that the tales about Fredericks cruelty to his own animals had been greatly exaggerated

By the evening of that day Napoleon was back at work, and on the next day it was learned that he had instructed Whymper to purchase in Willingdon some booklets on brewing and distilling.

A week later Napoleon gave orders that the small paddock beyond the orchard, which it had previously been intended to set aside as a grazing ground for animals who were past work, was to be ploughed up.

Figure 1. A sample

The sample above is a part of the summary of George Orwell's *Animal Farm*. As illustrated, the system identified the sentences that describe the development of the action. Expressions of time are quite obvious: *when time passed*, *a few days later*, *at the same time*, *by the evening*, and *a week later*. For validity reasons, however, a final test was carried out. A panel of judges was selected in order to see whether the automatic summaries are coherent and meaningful for readers and users. It was made sure that judges do not have a prior knowledge or background about the original texts. They were given the summaries to read and were asked some questions about the texts in order to see whether the information included in the summaries are meaningful and coherent. The judges were able to answer simple questions about the events in the original texts. For them, the extracted summaries were helpful for having a clear idea about the events in each story. They also indicated that summaries were not misleading. Based on the judges' responses, it can be claimed that extracted summaries are both meaningful and coherent.

7. Conclusion

The article discussed the issue of the difficulties of extracting automatic summarization of literary texts paying attention only to the problems of cohesion and coherence in the automatic summaries of prose fiction. It can be claimed that part of the problem is attributed to the peculiar nature of literary texts. Although texts of the kind have their own templates and have typical elements of plot, character, setting, etc, they are still considered as unstructured documents in terms of the automatic summarization theory. This entails that typical automatic summarization processes are not exclusively appropriate for processing literary data. This study proposed integrating LSA methods into automatic summarization processes so that extracted summaries are both meaningful and coherent. In spite of the success of the proposed system in addressing the problems of cohesion and coherence, some problems remain unresolved. One main problem is anaphora resolution. Within the extracted summaries, some problems concerning finding the right antecedents of the pronouns were identified.

Acknowledgments

This project was supported by the Deanship of Scientific Research at Prince Sattam Bin Abdulaziz University under the research project 4075/02/2015

References

Abramowicz, W. (2003). *Knowledge-based Information Retrieval and Filtering from the Web*. Boston; London: Kluwer Academic Publishers. https://doi.org/10.1007/978-1-4757-3739-4

Adrian, K., Stphane, D., & Tudor, G. (2007). Semantic clustering: Identifying topics in source code. *Information and Software Technology, 49*(3), 230-243. https://doi.org/10.1016/j.infsof.2006.10.017

Badry, R., Sharaf-eldin, A., & Elzanfally, D. (2013). Text Summarization within the Latent Semantic Analysis Framework: Comparative Study. *International Journal of Computer Applications, 81*(11).

Berry, M. W., & Castellanos, M. (2007). *Survey of Text Mining II: Clustering, Classification, and Retrieval*. London: Springer.

Bhatia, N., & Jaiswal, A. (2015). Trends in extractive and abstractive techniques in text summarization. *International Journal of Computer Applications, 117*(6). https://doi.org/10.5120/20559-2947

Bhatia, N., & Jaiswal, A. (2016). *Automatic text summarization and it's methods-a review.* Paper presented at the Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference. IEEE. https://doi.org/10.1109/CONFLUENCE.2016.7508049

Canhasi, E., & Kononenko, I. (2016). Weighted hierarchical archetypal analysis for multi-document summarization. *Computer Speech & Language, 37*, 24-46. https://doi.org/10.1016/j.csl.2015.11.004

Cha, J., & Kim, P. K. (2016). *The Automatic Text Summarization Using Semantic Relevance and Hierarchical Structure of Wordnet.* Paper presented at the International Conference on Broadband and Wireless Computing, Communication and Applications.

Chua, F., & Asur, S. (2013). Automatic Summarization of Events from Social Media. *ICWSM*.

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications.* Paper presented at the Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, 2002.

Cunningham, H., Wilks, Y., & Gaizauskas, R. J. (1999). *United Kingdom Patent No*. Sheffield: U. o. Sheffiled.

Deerwester, S., Susan, T. D., George, W. F., Thomas, K. L., & Richard, H. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391-407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). *Using latent semantic analysis to improve access to textual information.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, Washington, D.C., United States. https://doi.org/10.1145/57167.57214

Edmundson, H. P. (1969). New methods in automatic abstracting. *Journal of the Association for Computing Machinery, 16*(2), 264-285. https://doi.org/10.1145/321510.321519

Fattah, M. A. (2014). A hybrid machine learning model for multi-document summarization. *Applied intelligence, 40*(4), 592-600. https://doi.org/10.1007/s10489-013-0490-0

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge: Cambridge University Press.

Ferreira, R. (2013). Assessing sentence scoring techniques for extractive text summarization. *Expert systems with Applications, 40*(14), 5755-5764. https://doi.org/10.1016/j.eswa.2013.04.023

Fiori, A. (2014). *Innovative Document Summarization Techniques: Revolutionizing Knowledge Understanding*. New York: Information Science Reference. https://doi.org/10.4018/978-1-4666-5019-0

Foltz, P., Kintsch, W., & Landauer, T. K. (1998). Measurement of Text Coherence with Latent Semantic Analysis. *Discourse Processes, 25*, 285-307. https://doi.org/10.1080/01638539809545029

Gambhir, M., & Gupta, V. (2017). Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review, 47*(1), 1-66. https://doi.org/10.1007/s10462-016-9475-9

Gillespie, G. (1967). Novella, nouvelle, novella, short novel?—A review of terms. *Neophilologus, 51*(1), 117-127. https://doi.org/10.1007/BF01511303

Härdle, W., & Simar, L. (2003). *Applied multivariate statistical analysis*. Berlin; New York: Springer. https://doi.org/10.1007/978-3-662-05802-2

Hardy, T. (1878). *An Indiscretion in the Life of an Heiress*. London: The New Quarterly Magazine.

Hovy, E. (2005). Text Summarization. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (pp. 583-598). Oxford: Oxford University Press.

Hovy, E., Lin, C. Y., Mani, I., & Maybury, M. T. (1999). Automated Text Summarization in SUMMARIST. In *Advances in Automatic Text Summarization*. Cambridge: MIT Press.

Jackson, J. E. (1991). *A User's Guide to Principal Components*. New York: Wiley. https://doi.org/10.1002/0471725331

Jaradat, Y. A., & Al-Taani, A. T. (2016). *Hybrid-based Arabic single-document text summarization approach using genatic algorithm.* Paper presented at the Information and Communication Systems (ICICS), 2016 7th International Conference on. IEEE. https://doi.org/10.1109/IACS.2016.7476091

Jolliffe, I. T. (2002). *Principal Component Analysis*. Berlin; London: Springer.

Jones, S., Karen, M., & Maybury, M. (1999). Automatic Summarising: Factors and Directions. In *Advances in Automatic Text Summarization* (pp. 1-12). Cambridge: MIT Press.

Kazantseva, A., & Szpakowicz, S. (2014). *Measuring Lexical Cohesion: Beyond Word Repetition.* Paper presented at the Proceedings of the 25th International Conference on Computational Linguistics, Dublin, Ireland.

Kazantseva, A., & Szpakowicz., S. (2012). *Topical Segmentation: a Study of Human Performance and a New Measure of Quality.* Paper presented at the Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics. Proceedings (NAACL 2012), Montréal, Canada.

Kercheval, J. L. (1997). Short Shorts, Novellas, Novel-in-Stories. In *Building Fiction*. Cincinnati, Ohio: Story Press.

Ketui, N. (2014). *Thai Multi-document Summarization Based on Thai Elementary Discourse Units*: Sirindhorn International Institute of Technology, Thammasat University.

Knight, K., & Marcu, D. (2001). Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence, 13*.

Kogilavani, S. V., Kanimozhiselvi, C. S., & Malliga, S. (2016). Summary generation approaches based on semantic analysis for news documents. *Journal of Information Science, 42*(4), 465-476. https://doi.org/10.1177/0165551515594726

Landauer, T. K. (2007). *Handbook of latent semantic analysis*. Mahwah, N.J.; London: Lawrence Erlbaum Associates.

Landauer, T. K., Foltz, P., & Laham, D. (1998). An Introduction to Latent Semantic Analysis. *Discourse Processes, 25*(2-3), 259-284. https://doi.org/10.1080/01638539809545028

Lapata, M., & Barzilay, R. (2005). Automatic Evaluation of Text Coherence: Models and Representations. *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, 1085-1090.

Lee, C. B., Kim, M. S., & Park, H. R. (2003). Automatic Summarization Based on Principal Component Analysis. In F. M. P. a. S. Abreu (Ed.), *Progress in Artificial Intelligence* (pp. 409-413). Beja, Portugal. https://doi.org/10.1007/978-3-540-24580-3_46

Leibowitz, J. (1974). *Narrative Purpose in the Novella*. Michigan: University of Michigan Press. https://doi.org/10.1515/9783110883565

Li, C., Liu, F., Weng, F., & Liu, Y. (2013). Document Summarization via Guided Sentence Compression. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 490-500.

Li, W. (2015). *Abstractive Multi-document Summarization with Semantic Information Extraction.* Paper presented at the Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal. https://doi.org/10.18653/v1/D15-1219

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research Development, 2*(2), 159-165. https://doi.org/10.1147/rd.22.0159

Mani, I. (2001). *Automatic Summarization*. Amsterdam; [Great Britain]: Benjamins. https://doi.org/10.1075/nlp.3

Mani, I., & Maybury, M. T. (1999). *Advances in Automatic Text Summarization*. Cambridge, Mass; London: MIT

Press.

Mashechkin, I. V., Petrovskiy, M. I., Popov, D. S., & Tsarev, D. V. (2011). *Automatic text summarization using latent semantic analysis, 37*(6), 299-305.

Morariu, D. I. (2008). *Text Mining Methods Based on Support Vector Machine*. Bucure\1E63ti: Matrix Rom.

Moreno, L. (2016). *Software documentation through automatic summarization of source code artifacts.* Unpublished PhD, The University of Texas, Dallas.

Nenkova, A., & McKeown, K. (2011). Automatic Summarization. *Foundations and Trends in Information Retrieval, 5*(2-3), 103-233. https://doi.org/10.1561/1500000015

Ozsoy, M. G., Cicekli, I., & Alpaslan, F. N. (2010). Text Summarization of Turkish Texts using Latent Semantic Analysis. *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010),* 869-887.

Patil, A., Pharande, K., Nale, D., & Agrawal, R. (2015). Automatic Text Summarization. *International Journal of Computer Applications, 109*(17). https://doi.org/10.5120/19418-0910

Poibeau, T. (2013). *Multi-source, Multilingual Information Extraction and Summarization*. Berlin: Springer. https://doi.org/10.1007/978-3-642-28569-1

Porzel, R. (2010). *Contextual Computing: Models and Applications*. Berlin: Springer Science & Business Media.

Rico-Jimenez, J. J., Campos-Delgado, D. U., Villiger, M., & Otsuka, K. (2016). Automatic classification of atherosclerotic plaques imaged with intravascular OCT. *Biomedical Optics Express, 7*(10), 4069-4085. https://doi.org/10.1364/BOE.7.004069

Robert, L. D., Kevin, W. D., & Laura, A. M. (2000). *A comparison of rankings produced by summarization evaluation measures*. Paper presented at the Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization.

Rutkauskas, Ž., & Bargelis, A. (2016). Knowledge–based method for gate and cold runner definition in injection mold design. *Mechanics, 66*(4).

Saggion, H., & Poibeau, T. (2012). Automatic Text Summarization: Past, Present and Future. In R. Y. T. Poibeau, H. Saggion, & J. Piskorski (Ed.), *Multi-source, Multilingual Information Extraction and Summarization* (pp. 3-13). Berlin: Springer.

Salton, G. (1997). Automatic text structuring and summarization. *Information Processing & Management, 33*, 193-207. https://doi.org/10.1016/S0306-4573(96)00062-3

Salton, G., & Buckley, C. (1987). *Term Weighting Approaches in Automatic Text Retrieval*. Cornell: Cornell University.

Salton, G., & Buckley, C. (1988). Term-weighing approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513-523. https://doi.org/10.1016/0306-4573(88)90021-0

Shah, C., & Jivani, A. (2016). *Literature Study on Multi-document Text Summarization Techniques.* Paper presented at the International Conference on Smart Trends for Information Technology and Computer Communications. https://doi.org/10.1007/978-981-10-3433-6_53

Silva, G. (2015). Automatic text document summarization based on machine learning. *Proceedings of the 2015 ACM Symposium on Document Engineering. ACM*. https://doi.org/10.1145/2682571.2797099

Snowsill, T. (2012). *Data mining in text streams using suffix trees.* Unpublished Thesis (Ph.D.), University of Bristol.

Srivastava, A. N., & Sahami, M. (2009). *Text mining: Classification, Clustering, and Applications*. London: Chapman & Hall/CRC. https://doi.org/10.1201/9781420059458

Tan, H. (2012). *Knowledge Discovery and Data Mining*. Berlin: Springer Science & Business Media. https://doi.org/10.1007/978-3-642-27708-5

Tonfoni, G. E. (1985). *Artificial intelligence and text-understanding: plot units and summarization procedures*. Parma, Italy: Edizioni Zara.

Torres-Moreno, & Juan-Manuel. (2014). *Automatic text summarization*. Hoboken, New Jersey: John Wiley & Sons. https://doi.org/10.1002/9781119004752

Torres-Moreno, J. M. (2014). *Automatic Text Summarization*. Hoboken, New Jersey: John Wiley & Sons.

Wang, S., Li, W., Wang, F., & Deng, H. (2010). A Survey on Automatic Summarization. *International Forum on Information Technology and Applications*, 193-196. https://doi.org/10.1109/ifita.2010.96

Yeh, J. Y. (2005). Text summarization using a trainable summarizer and latent semantic analysis. *Information Processing & Management, 41*(1), 75-95. https://doi.org/10.1016/j.ipm.2004.04.003

Zanasi, A., Brebbia, C. A., & Ebecken, N. F. F. (2005). *Data Mining VI: Data Mining, Text Mining and their Business Applications*. Southampton: WIT.

Zhuge, H. (2016). *Multi-Dimensional Summarization in Cyber-Physical Society*. Amsterdam: Elsevier.