A PATTERN-GROWTH SENTENCE COMPRESSION TECHNIQUE FOR MALAY TEXT SUMMARIZER

SURAYA ALIAS

UNIVERSITI SAINS MALAYSIA
2018

A PATTERN-GROWTH SENTENCE COMPRESSION TECHNIQUE FOR MALAY TEXT SUMMARIZER

by

SURAYA ALIAS

**Thesis submitted in fulfillment of the requirements
for the degree of
Doctor of Philosophy**

January 2018

ACKNOWLEDGEMENT

In the name of Allah, the Most Gracious and the Most Merciful.

I would like to extend my deepest praise to Allah S.W.T for the blessings and guidance in completing this research and thesis. Also special thanks and gratitude to;

The Ministry of Higher Learning and Universiti Malaysia Sabah for the financial support throughout my studies, and also Universiti Sains Malaysia for the resources and facilities provided.

My supervisors, Dr Siti Khaotijah Mohammad and Dr Tan Tien Ping, for their kind supervision and knowledge in guiding me throughout this research.

My parents, Allahyarham Hj Alias bin Kassim and Hajah Samsiah bt Zainal, for being supportive and always being there when I needed most.

My husband, Mohd Shamrie bin Sainin, my sons, Aiman Haziq and Aish Hafiy, and daughter Damia Hana, for the love, patience, trust and inspiration they had given me.

And lastly;
My friends and others, this research will not be completed without you.

At Taubah:129 "Cukuplah Tuhan bagiku, tiada Tuhan selain dariNya. Hanya kepadaNya aku bertawakal..."

This thesis is dedicated to all my beloved, for you, I give my all.

TABLE OF CONTENTS

CHAPTER 1 – INTRODUCTION

CHAPTER 2 – LITERATURE REVIEW

CHAPTER 3 – RESEARCH METHODOLOGY

CHAPTER 4 – PATTERN-GROWTH TEXT REPRESENTATION MODEL

CHAPTER 5 – HUMAN COMPRESSION PATTERN

CHAPTER 6 – MALAY TEXT SUMMARIZER MODEL

CHAPTER 7 – SUMMARY EVALUATION AND DISCUSSION

CHAPTER 8 – CONCLUSION AND FUTURE WORKS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

AR	Association Rule
ATS	Automatic Text Summarization
BLIS	Bernama Library & Infolink Service
BOW	Bag-of-Words
C&L	(Clarke & Lapata, 2008)
CLP	Computational Linguistics
CP	Closed Patterns
CR	Compression Ratio
DUC	Document Understanding Conference
EBMT	Example-Based Machine Translation
FASP	Frequent Adjacent Sequential Pattern
FASPe	Frequent Eliminated Pattern
FP	Frequent Patterns
FSP	Frequent Sequential Patterns
ILP	Integer Linear Programming
IR	Information Retrieval
K&M	(Knight and Marcu, 2002)
KSMI	Kamus Multimedia Bahasa Melayu - Bahasa Inggeris
LSA	Latent Semantic Analysis
LSI	Latent Semantic Indexing
MFS	Maximal Frequent Patterns
MMR	Maximal Marginal Relevance
MYTextSum	Malay Automatic Text Summarization model
NER	Named Entity Recognition
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
PGSC	Pattern-Growth Sentence Compression
POS	Part-of-Speech
PTM	Pattern Taxonomy Model
PW	Penanda Wacana (discourse marker)
ROUGE	Recall-Oriented Understanding for Gisting Evaluation
RR	Retention Ratio
RST	Rhetorical Structure Theory
SC	Sentence Compression
SPM	Sequential Pattern Mining
SVD	Singular Value Decomposition
TAC	Text Analysis Conference
VSM	Vector Space Model

LIST OF SYMBOLS

$\subseteq$	Is a subset (contained in), subsequence
$\in$	Is an element of
$\rightarrow$	Implies
$\cap$	AND, JOIN
$\cup$	OR, UNION
α	*prefixTerms*
β	Next *adjacent sequence* term or $t_{(k+1)}$
σ	*min_sup*
δ	*min_conf*
$\{\quad\}$	Set
$=$	Equals
$==$	Equality operator
$\equiv$	Equivalence
$\forall$	For all
$\geq$	Greater than or equal to
$\leq$	Less than or equal to
$>$	Greater than
$<$	Less than
$\sum$	Summation - sum of all values in range of series
$[1 \dots n]$	Closed interval

TEKNIK PEMAMPATAN AYAT MENGGUNAKAN PERTUMBUHAN CORAK UNTUK PERINGKASAN TEKS BAHASA MELAYU

ABSTRAK

Peringkasan teks secara automatik (ATS) telah memberi manfaat kepada pengguna dengan membantu dari segi mengenal pasti dan mengekstrak maklumat yang penting dari teks tertentu dengan lebih mudah. Tujuan aplikasi teknik Pemampatan Ayat (SC) di dalam bidang ATS adalah untuk menggugurkan unsur yang tidak penting dalam sesebuah ayat di dalam ringkasan di samping mengekalkan unsur yang penting dengan mengekalkan tatabahasanya agar ayat tersebut tidak terjejas. Kebanyakan teknik SC yang terdahulu mempunyai pergantungan yang tinggi kepada peraturan sintaktik dan pengetahuan pada perkataan individu atau frasa ayat untuk proses pengguguran unsur. Walaupun ianya mampu menghasilkan ayat mampat yang mematuhi tatabahasa, pendekatan sebelum ini masih mempunyai beberapa kelemahan seperti kegagalan untuk memasukkan beberapa ayat yang penting dan relevan dalam pembinaan sesebuah ringkasan akhir. Kajian ini menumpukan kepada penemuan corak mampatan manusia dari korpus ringkasan bahasa Melayu yang dibangunkan untuk meningkatkan kebolehbacaan dan keberkesanan maklumat ringkasan yang dihasilkan. Satu teknik baru pemampatan ayat menggunakan pertumbuhan corak (PGSC) yang diilhamkan menggunakan strategi "pecah dan perintah" untuk bahasa Melayu dicadangkan di dalam tesis ini. Idea dasarnya adalah untuk membahagikan ayat-ayat kepada segmen, di mana segmen-segmen yang tidak penting digugurkan sementara segmen yang penting ditakluk secara berulang. Satu perwakilan teks baru berdasarkan corak dengan “kekangan corak teks" yang ditemui di dalam kajian ini berfungsi untuk mengenal pasti maklumat penting daripada dokumen teks. Sementara itu, satu set Peraturan Pengguguran Ayat dengan nilai keyakinan *Conf* telah ditemui dari corak pemampatan ayat para panel bahasa, di mana ia berfungsi untuk menunjukkan unsur yang sering digugurkan di dalam ayat. Keputusan pengguguran dalam teknik PGSC ini adalah hasil gabungan kedua-dua corak teks yang ditemui yang memenuhi "kekangan penyingkiran" yang dicadangkan. Eksperimen yang dijalankan telah menampakkan kejayaan di mana ringkasan termampat melaporkan nilai F-Measure sebanyak 0.5752 apabila dibandingkan dengan ringkasan yang dihasilkan oleh panel bahasa, dan ia juga mengatasi kaedah baseline (ringkasan yang tidak termampat). Penilaian manual telah menghasilkan nilai purata kebolehbacaan 4.31 daripada 5, dan 4.1 untuk kandungan responsif, di mana ianya menunjukkan kualiti yang lebih baik dan kebolehbacaan bagi ringkasan termampat yang dihasilkan oleh model yang dicadangkan.

A PATTERN-GROWTH SENTENCE COMPRESSION TECHNIQUE FOR MALAY TEXT SUMMARIZER

ABSTRACT

Automatic Text Summarization (ATS) has benefited users in terms of identifying and extracting the most salient information from a given text with less effort. The application of Sentence Compression (SC) in ATS is to remove unimportant constituents from a summary sentence while preserving the salient ones by keeping the sentence's grammar intact. Most previous SC techniques have a high dependency on syntactic rules and knowledge applied to individual word or phrase to cater the removal decision. Despite the ability to produce a new grammatical compressed sentence, prior approaches still suffer several drawbacks including the failure to include some significant and relevant sentences in constructing the final summary sentence. This study focuses on discovering human compression pattern from the developed Malay summary corpus to improve the readability and informativeness of the produced summary. A new Pattern-Growth SC (PGSC) technique inspired by the "divide and conquer" strategy tailored to the Malay language is proposed. The underlying idea is to divide the sentences into segments where unimportant segments are removed while the important ones are conquered iteratively. A new pattern-based representation with "textual constraints" discovered in this study serves as a feature to identify significant information from the text document. Meanwhile, a set of Sentence Elimination Rules with confidence value *Conf* discovered from human compression pattern indicates the constituents that are frequently removed. The removal decision is based on both discovered textual patterns fulfilling the proposed “removal constraints”. The experiments have shown promising results where the compressed summaries reported an F-Measure score of 0.5752 when compared to the gold standard human summaries and perform better than the baseline (uncompressed) methods. Manual evaluation produced average readability score of 4.31 out of 5 and 4.1 for content responsiveness, suggesting a better quality and readability of the compressed summaries produced by the proposed model.

CHAPTER 1

INTRODUCTION

1.1 Introduction

In this digitalized era, users are overwhelmed with the vast amount of information that is made available online. In order to reduce the time for searching the important facts from the overloaded information, a summary can provide an insight of related information with less effort. Automatic Text Summarization (ATS) is an interdisciplinary research area related to Information Retrieval (IR), Natural Language Processing (NLP) and Computational Linguistics (CL). It is an automated process of creating a summary from a single or multiple document input sources. The output type of summary can either be generic, query-focused (based on specific user-topic), or sentiment-based such as summarizing user's opinion. Meanwhile, the methods to produce an automated summary can be performed via extractive or abstractive methods (Das & Martins, 2007; Hahn & Mani, 2000).

An extractive method selects and concatenates the most important sentences to produce a shorter version of a document. Here, sentences are scored by its importance based on certain features such as surface and content. Among the common surface level features are the title, sentence's position and word frequency. Meanwhile, the content feature refers to sentences that carry the most significant information and contains the topic words in the document (Edmundson, 1969; Ferreira et al., 2014; Litvak & Last, 2013; Luhn, 1958; Wong, Wu, & Li, 2008). To produce the summary, the extractive summarizer selects the highest score and most representative sentence without modifying the original source sentence.

In contrast, an abstractive method constructs a summary by modifying, paraphrasing and joining related information to form a new sentence. This method mimics the human-made summary where extensive Natural Language Processing (NLP) and prior knowledge is needed. Due to the complexity in generating an abstractive based summary, the extractive summarization method has dominated the research area in the ATS field until today where the issue on producing a quality automated summary still opens for improvement as discussed in Gambhir and Gupta (2017) and Fang, Mu, Deng, and Wu (2017).

In ATS, to identify the main topic from the input text, most extractive summarizer models employ the following three tasks in generating an extractive summary as described in Nenkova and McKeown (2012). Firstly, an intermediate representation of the input content is created, which contains the key source or the main topic of the text. The input can be represented using a list of features such as a vector of words, N-grams and graphs model with different levels of granularity (term, sentence or document level). Next, the sentence scoring task is performed based on the respective representation. The summarizer model employs various methods such as statistical, machine learning and graph model to estimate the relevance and importance of each extracted sentence. Finally, sentence selection is done, which refers to the process of generating a summary. The summarizer should decide based on the length of summary and certain sentence selection techniques such as greedy approach, global optimization algorithm and clustering approaches, which sentence should be selected by considering the coverage and its importance.

Nevertheless, although automated summary generation using sentence extraction method can provide users with sufficient information; during the process, the extracted sentence might contain both essential and also extraneous information in the same sentence since the content is directly copied (verbatim). This extraneous constituent may consist of any granularity; a single word (term) or even a phrase. Thus, including this extraneous or unnecessary constituent because it happens to be in the same sentence that bears important facts may have an effect on the readability and coherence of the summary (Perera & Kosseim, 2014; Saggion & Poibeau, 2013).

This underlying issue has sparked interest in the ATS research community with one of the proposed solutions is using Sentence Compression (SC). Jing (2000) defined SC as an independent task or problem in ATS where: a) unimportant details from a sentence are eliminated, b) salient information is preserved, and c) the sentence grammar is kept intacted. SC can also be viewed as a scaled down version of summarization performed at a sentence level where the problem is typically formulated as a word deletion task (Knight & Marcu, 2000, 2002). The compressed sentence is constructed by removing tokens from the source sentence without applying any paraphrasing or reordering operation such as in abstractive method. Some leading researchers in this field (Cohn & Lapata, 2008; Galanis & Androutsopoulos, 2010) defined this deletion-based approach as an extractive compression differentiating between extractive and abstractive approaches in ATS. The SC approach has been primarily used in single document summarization (Jing, 2000; Knight & Marcu, 2002; Turner & Charniak, 2005), which has been currently applied in the multi-document summarization area later on by (Boudin & Morin, 2013; Filippova, 2010; ShafieiBavani, Ebrahimi, Wong, & Chen, 2016; Wang, Raghavan, Castelli, Florian, & Cardie, 2013).

Apart from summarization, sentence compression technique also benefited other applications including producing TV headlines (Dorr, Zajic, & Schwartz, 2003) and subtitles (Vandeghinste & Pan, 2004). SC also has been used to assist impaired citizens in Grefenstette (1998) and to produce more compact sentences for smaller screen such as in phones and personal digital assistant (PDA) previously by Corston-Oliver (2001). With limited space available, it seems practical to produce a compressed sentence rather than displaying a full extracted sentence for the user to view.

Table 1.1 illustrates an example of abstractive and extractive compression performed on a single sentence in producing an automated summary. The abstractive compression involves some word re-ordering and paraphrasing activities such as replacing the word “closed” to “halted”. Meanwhile, in extractive compression, some words are removed without changing the word order to preserve only the most useful and significant facts to the user.

Table 1.1: An example of abstractive and extractive sentence compression from (Clarke, 2008; Thadani & McKeown, 2013).

Sentence 1	***Production was closed down at Ford last night for the Christmas period***
Sentence 1-a (*abstractive* compression)	*Ford production was halted yesterday for the holidays*
Sentence 1-b (*extractive* compression)	*Production closed at Ford for Christmas*

1.2 Problem Background

The direction of this study is in the area of Automatic Text Summarization using extractive method to summarize a single document by focusing on developing a new extractive sentence compression technique for Malay language. To generate a summary, many developed summarizers model employed the commonly used text representation using language models such as the traditional Bag-of-Words (BOW) to represent each term in a text as an *n*-dimensional vector of keywords such as in

(Conroy, Schlesinger, O'leary, & Goldstein, 2006; Erkan & Radev, 2004; McDonald, R., 2007). Meanwhile, others have exploited the N-gram probabilistic language model as sentence features (Clarke & Lapata, 2008; Matsuo & Ishizuka, 2004) and representing it using a graph model (Ganesan, Zhai, & Han, 2010; Mihalcea & Tarau, 2004). Nevertheless, there are some known issues in these language models such as inaccurate semantic representation and misleading meanings in BOW, especially when handling similarity problems in a sentence as the word order is not preserved (Ning, Yuefeng, & Sheng-Tang, 2012). On the other hand, a known issue for the N-gram model is a high dimensionality of word size combination where not all combinations are available across the collection, which is identified as the data sparsity issue (Kim, Park, Lu, & Zhai, 2012; Le & Mikolov, 2014).

Due to this, some researchers have shifted towards manipulating the Frequent Pattern (FP) found in text or "textual patterns" by proposing a pattern-based summarizer model. The pattern-based representation with the ability to correlate between words can provide a natural text representation while preserving the word's semantics relationships. For example, Ledeneva, Gelbukh, and García-Hernández (2008) experimented the use of Maximal Frequent Patterns (MFS) to represent the significant content from a document. Their pattern-based summarizer model has resulted in improvement in single extractive summary Recall value in comparison to the BOW and N-gram language model. Meanwhile, current attempt by Qiang, Chen, Ding, Xie, and Wu (2016) using Closed Patterns (CP) representation to remove redundant sentences and preserve the important ones have demonstrated the use of pattern-based summarizer model in the area of multi-document summarization. On top other that, a recent work by Xie, Wu, and Zhu (2017) propose the use of wildcards constraints to extract Frequent Sequential Patterns (FSP) as the key phrase that identifies the important topic from a document. Their work shows promising results using a benchmark key phrase dataset, which indicates the viability of the pattern-based representation.

According to Gupta and Han (2011), the concept of Sequential Pattern Mining (SPM) pioneered by Agrawal and Srikant (1995) in transactional database can be applied to discover regularities (patterns) in text data since a sentence can be viewed as a sequence of items or words. SPM is generally categorized into two (2) methods namely Apriori-based and Pattern-Growth approaches. The Pattern-Growth approach implements the "divide-and-conquer" strategy benefitting small search space in data structure with reduced candidate generation cost compared to the "generate-and-test" strategy in Apriori-based method (Aggarwal, C. C. & Han, 2014; Pei et al., 2004).

At the time this research started, there is no readily available gold-standard summaries dataset in Malay language. A gold-standard summary is a human crafted summary that is used to evaluate the performance of the summary produced by the summarizer model. Most prior and recent studies in extractive text summarization are commonly in English language. The available benchmark dataset is also mainly in English language particularly from the Document Understanding Conference (DUC)[1] and Text Analysis Conference (TAC)[2] organized by the National Institute of Standards and Technology (NIST). However, it is refreshing to know that there are ongoing efforts nowadays in other languages through the development of their own corpora such as in Arabic (Belkebir & Guessoum, 2015), Portuguese (Nóbrega & Pardo, 2016) and Vietnamese (Thu, Ngoc, Ngoc, & Huynh, 2016). This progress has shown positive impact to the effort of preserving their respective native language, which encouraged

[1] http://duc.nist.gov/
[2] http://tac.nist.gov/

this current study to generate a new benchmark dataset in the area of SC and ATS since there is no formal baseline SC technique for extractive summarizer model for the Malay language.

Literature in ATS focusing on SC techniques has been an interest to researchers as a way to improve the quality of extractive summary produced. Some known SC techniques such as linguistics rule based (Conroy et al., 2006; Jing, 2000; Jing & McKeown, 1999; Zajic, Dorr, Lin, & Schwartz, 2007), statistical (Galley & McKeown, 2007; Knight & Marcu, 2000, 2002), machine learning (Nguyen, Phan, Horiguchi, & Shimazu, 2007; Turner & Charniak, 2005), keyword-based (Conroy et al., 2006; Prasad Pingali & Varma, 2007) and integer linear programming (ILP) by (Clarke & Lapata, 2008; Cohn & Lapata, 2008) have been previously explored. Furthermore, some recent works in this area also include graph related optimization by (Boudin & Morin, 2013; Filippova, 2010; Filippova & Strube, 2008) that has been applied in the area of multi-sentence compression.

Most of the aforementioned SC techniques perform extractive compression and are highly dependent on syntactic knowledge (syntactic parser and dependency parser) applied to individual word or phrases to decide on the compressions decision. The main reason for this dependency is to avoid composing ungrammatical sentence after the compression process. Thus, it is common for a summarizer model to perform the compression by referring to syntactic tree such as in (Filippova & Strube, 2008; Galley & McKeown, 2007; Knight & Marcu, 2002) and referring to the sentence's global discourse information by Clarke and Lapata (2008). Nonetheless, there is still trade-off that exists between the model's performance in balancing both grammatical and informativeness of the summary (Katja, Enrique, Carlos, Lukasz, & Oriol, 2015).

An empirical study by Lin (2003) claimed that even though pure syntactic-based compression approach has performed grammatically well, it has insignificant improvements on the summary's content evaluation using the DUC 2001 dataset. Similar results were also found in current experiments conducted by Perera and Kosseim (2014), where the syntax-driven compression approach gave the lowest content agreement with human summaries using DUC 2007 dataset.

Thus, recently, more researchers have raised their concern on the syntactic dependency issue as discussed in Katja et al. (2015) and Thu et al. (2016). This happens following an experimental attempt by Filippova (2010) and then followed by Boudin and Morin (2013) that uses only minimum Part-of-Speech (POS) tagging and list of stop words to find the shortest paths in word graphs for sentence compression. Fillipova's straightforward attempt that simply relies on the words of the sentences has shed some lights to the viability of being less dependent on the syntactic approach. However, the experimental approach still failed short in preserving salient information to be added in the final summary generation. Thus, the syntactic dependency problem is yet isolated and open for improvement in SC research area.

In the application of pure syntactic-based SC technique to the Malay language, limited Malay NLP tools and resources have become among the hurdles and challenges. For example, the POS tagger and parser (Alfred, Mujat, & Obit, 2013; Mohamed, Omar, & Ab Aziz, 2011; Xian et al., 2016) and Named Entity Recognition (NER) (Alfred, Leong, On, & Anthony, 2014; Zamin & Bakar, 2015) tools for Malay language are not yet publicly available since they are still actively studied (Alfred et al., 2013). However, this challenge becomes the motivation for this study to explore alternative approach rather than depending on NLP tools and syntactic approaches by investigating a new pattern-based extractive SC technique.

1.3 Statement of Problem

In extractive summarization, the challenge is to determine which units of text such as sentences, phrases or paragraph is important to be extracted and selected to generate a summary. In achieving this, the application of sentence compression to eliminate unimportant constituents extracted together while performing the sentence selection is investigated. Following are the issues stated in this study.

Firstly, this study highlights the problems in identifying and representing significant information in an extractive summary. This is because the basis of the SC task is not only to remove unnecessary constituent in a sentence, but also to preserve the significant ones. Existing language model such as the BOW model has a known limitation, which is an inaccurate semantic representation of text since the word order is not preserved. Meanwhile, the N-gram model comes with high dimensionality of word size combination where not all grams are meaningful to be used in representing the text (Kim et al., 2012; Le & Mikolov, 2014).

Secondly, the problem of non-existence gold-standard summaries dataset to evaluate the produced summary and to discover humans' compression pattern for Malay language. The nearest ATS work in Malay language related to this study is from Jusoh, Masoud, and Alfawareh (2011) and Zamin and Ghani (2011). The dataset used by Zamin and Ghani (2011) only consists of small samples using 10 Malay articles with no application of SC. Meanwhile, Jusoh et al. (2011) tried to directly refine a summary sentence using a static list of eliminated words translated from English using the sample of 40 Malay articles. However, they only reported their result in terms of summary compression rate, whereas the effects of their refinement technique on the grammatical and content informativeness of the summary were not discussed and evaluated. Hence, the initiation of a new Malay summary corpus for this study is essential since there is no formal baseline summarizer model available for Malay language. Besides, this study aimed to discover the human's sentence compression pattern from the corpus in developing the new SC technique.

During the employment of this study, even though the existing benchmark English news corpus data from DUC and TAC conferences can be directly translated into Malay language using available translation tools such as Google Translate[3] and Example-Based Machine Translation (EBMT), the translated sentences still needs to be aligned and manually validated especially when handling complex sentences, which incurs delicate and expensive labour work as experienced in Kwee, Tsai, and Tang (2009). Moreover, since the grammar for each compressed sentence needs to be validated upon the respective language used by the summarizer model, it is difficult to generalize the syntactic transformation process from one language to another.

Thirdly, the following problem is that most traditional methods in Sentence Compression are highly dependent on syntactic knowledge such as referring to syntactic parser and dependency parser to decide on the compressions decision of removing unimportant constituent from a sentence (Almeida & Martins, 2013; Filippova & Strube, 2008; Gagnon & Da Sylva, 2006). Their methods heavily rely on external knowledge resources such as WordNet, lexicon database and sentence's discourse information incur some processing cost. Nonetheless, despite elegantly producing a new grammatical sentence, the prior approach still suffer some drawbacks such as missing some significant and relevant sentences in constructing the final summary sentence (Boudin & Morin, 2013).

[3] https://translate.google.com/

It was found that during the summary evaluation experimented in Lin (2003), the syntactic methods do not necessary improve the overall summary content. This is due to some important content might have been deleted since the method refers to syntactic token deletion. The finding is also supported by Perera and Kosseim (2014) where their experiment results show syntactic compression methods such as syntactic with relevancy, relevancy-driven and syntax-driven gave the lowest content agreement when evaluated against gold-standard human summaries. Thus, a more subtle compression such as the keyword-based method that refers to human compression pattern that shows promising results is worth for further investigation.

1.4 Research Questions

The questions this study attempts to answer are:

1) Is the proposed pattern-based representation viable to be used as text features to represent significant information from text documents as compared to existing language models?

2) How do the rules from human compression pattern assist the proposed Sentence Compression technique for Malay ATS model?

3) Can the proposed Pattern-Growth Sentence Compression technique improve the readability and informativeness of a Malay text summary?

1.5 Research Objectives

In general, this study aims to develop an Automatic Text Summarizer model using extractive method in summarizing a single news article. Developing this model involved examining the effects of applying Sentence Compression technique on the Malay language.

The specific objectives of this study include:

1) To propose a new pattern-based text representation model based on the Pattern-Growth technique as text features to represent significant information in a text document.
2) To discover human compression pattern by initiating a gold-standard Malay summary corpus, where the corpus is to evaluate the performance of the proposed Malay Text Summarizer model.
3) To improve the quality of an automated Malay text summary by proposing a new Pattern-Growth Sentence Compression technique tailored for the Malay language.

The hypothesis of the study is stated as below:

"*The application of Sentence Compression technique on the Malay language can improve the quality of summary produced by the Automatic Text Summarization model*".

1.6 Research Scope

- The study focused on extractive text summarization method in summarizing a single document input text. The total of 100 Malay news articles covering the Natural Disaster (ND) and Events topics in Malaysia for a specific period of time was downloaded and used for experiments by following the English DUC 2002 dataset preparation (Appendix E).
- This study produced an ATS model that employs extractive SC technique as one of the tasks where it is tailored to Malay language. Since there is no baseline summarizer model with SC technique formally developed in Malay language, the study did not perform any comparison with existing SC technique implemented in English or other languages since each language has their own grammar pattern. The summaries produced by the model in this study are evaluated against the gold-standard summaries produced by the panelists.
- The performance of the proposed model is evaluated using the Recall-Oriented Understanding for Gisting Evaluation (ROUGE) toolkit developed by Lin (2004b), which is a benchmark multilingual automatic evaluation tool based on the metrics of Recall, Precision and F-measure values against human summaries (gold standard). Human evaluation by Malay language experts is also performed in this study based on the evaluation method used in DUC 2005[4].

[4] http://www-nlpir.nist.gov/projects/duc/guidelines/2005.html

1.7 Expected Contribution

This study is expected to contribute in:

1) **Initiating a new Malay summary corpus** consisting gold-standard summaries and samples of sentence compression data for the use in the research area of ATS and SC.

2) **Developing a new pattern-based text representation model** to identify significant information in a text.

3) **Discovering a new set of human compression pattern** for Malay language where it represents a set of words or phrases frequently eliminated by human summarizers when composing a summary.

4) **Introducing a new Pattern-Growth Sentence Compression** (PGSC) technique inspired by the "divide-and-conquer" approach tailored to Malay language in the area of extractive sentence compression.

1.8 Organization of the Thesis

The organization of this study is as follows:

Chapter 2 reviews related literature by focusing on processes involved in producing an extractive summary, corpus development and state of the art technique in sentence compression. On top of that, the basis of SPM Pattern-Growth technique that is referred in this study is described in this chapter.

Chapter 3 describes the methodology used in this study. It includes the description of English DUC 2002 benchmark data and new data for the Malay summary corpus. Furthermore, it describes the working framework of the proposed Malay ATS model and evaluation metrics used in this study.

In Chapter 4, the development of the proposed pattern-based text representation model is described, whereas Chapter 5 presents the details on Malay summary corpus development and analysis. From the analysis, a set of compression pattern is discovered where the Sentence Elimination Rules comes with *Conf* value.

Chapter 6 consolidates the development work of the new PGSC technique proposed in this study. This chapter demonstrates the combination of pattern-based text representation and the discovered compression pattern together with removal constraints in conquering only significant segments in a sentence.

Chapter 7 provides the discussion on experiments and findings from the application of SC in this study. Automatic and manual evaluation is performed on the extractive summary produced by the proposed summarizer model.

Finally, Chapter 8 presents the conclusion from the study followed by main findings derived from the analysis of the results produced. Future work and improvement are also stated here.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In this digital era, managing and condensing abundant information available online has necessitated the need for ongoing study in Automatic Text Summarization (ATS). Much advancement has been made since the first seminal work in ATS during the late 1950s by Luhn (1958). To date, the application of text summaries is to commercially cater to the needs of user's personal gadgets such as mobile devices and tablets. Specifically, all works involved in ATS focuses on finding a way to bridge the gap between human-made summaries and the automated ones. For this reason, the need to improve the quality of an extractive summary by applying SC technique has become the goal of this study.

The organization of this chapter is as follows: In Section 2.2, the background of an Automatic Text Summarizer system framework is presented. Meanwhile, Section 2.3 presents the existing text summarization technique. In Section 2.4, the Topic Identification process in ATS which include: 1) creating an intermediate representation of the input text, 2) sentence scoring and 3) summary sentence selection described in (Nenkova & McKeown, 2011, 2012) is detailed out. Then, in Section 2.5, the basis of the proposed Pattern-Growth Sentence Compression (PGSC) technique by reviewing the SPM technique is demonstrated. Section 2.6 presents existing work in sentence compression, while Section 2.7 provides the discussion to identify the gap and alternatives that can be explored rather than the traditional syntactic dependency approach.

2.2 Background of Automatic Text Summarization

The first work in Automatic Text Summarization (ATS) involves summarizing a single technical English document by Luhn (1958). From there, vast enhancements have been done where the issue in producing a quality automated summary is still open for research study (Gambhir & Gupta, 2017).

The framework of an ATS model involves three general stages as illustrated in Figure 2.1. It consists of 1) Topic Identification, 2) Tansformation and 3) Summary Generation (Hahn & Mani, 2000; Hovy, 2005; Jones, K. Sparck, 1999; Radev, Dragomir R., Hovy, & McKeown, 2002).

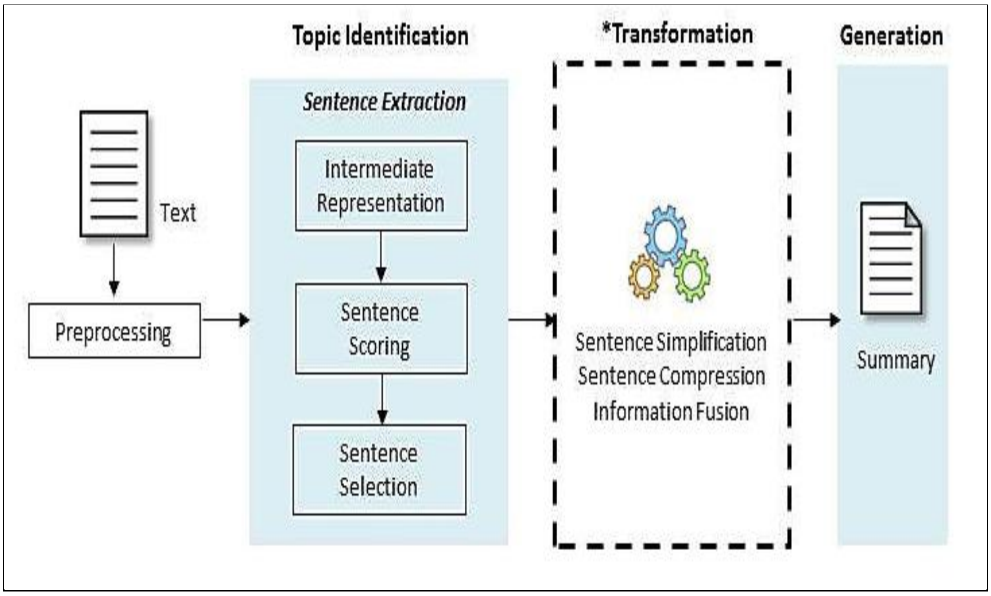

Figure 2.1: General framework of an ATS model

Both extractive and abstractive summarizers usually perform the topic identification stage. In this initial stage, the summarizers should identify which portion of the original source should be selected and included in the summary based on the identified topic**.** To do this, most summarizers will include three main tasks that include creating an intermediate representation of the input text, sentence scoring and sentence selection.

Next, in the transformation or interpretation stage, the selected sentence undergoes certain transformation using methods such as sentence simplification, sentence compression, and information fusion technique (Nenkova & McKeown, 2011). Abstractive summarizer usually caters this stage since it involves deep NLP knowledge in interpreting the original sentence before transforming it into a new summary sentence. In practice, a full extractive summarizer model that only extracts or copies the original sentence would normally skip this transformation stage (Hovy, 2005).

The sentence simplification technique simplifies the sentence structure and opts for simpler word choice to reduce the length of sentence (Finegan - Dollak & Radev, 2015). For example, rewriting the passive phrase with active ones and using simpler reference for noun and pronouns. This approach has been previously used to assist users with linguistic disabilities including the blind (Grefenstette, 1998). Meanwhile, information fusion deals with combining and restructuring the pieces of information from sentences together and removing unnecessary ones where the output sentence is more towards abstractive manner (Barzilay & McKeown, 2005; Barzilay, McKeown, & Elhadad, 1999). This is why sentence compression technique with the aim of removing unimportant details on a sentence level and preserving the important ones has become the goal for current researchers to improve the quality of an extractive summary.

Finally, in the ATS framework, the summary generation stage is based on the requirement and formatting such as preparing a summary for mobile devices or newspaper headlines. The work in this study is towards extractive summarization to produce a single extractive summary of generic news. On top of that, the effects of applying sentence compression task were also investigated.

2.3 Extractive Text Summarization Techniques

This section briefly introduces existing techniques applied in extractive summarization, which are generally classified into statistical (feature-based), machine learning, semantic and discourse approach and graph model. The pattern-based approach referred in this study is discussed in Section 2.5 with the introduction to Sequential Pattern Mining (SPM) technique.

2.3.1 Statistical-based

A statistical-based summarizer model exploits the features from documents to extract important sentences to be added in a summary. The higher the sentence score, the higher the chance for a sentence to be chosen. Luhn (1958) is known as the pioneer in ATS area using this statistical approach. His work was based on word frequency and phrases by focusing on technical documents. A decade later, some common surface level features have been used by Edmundson (1969) to mark the importance of a passage or sentences such as title, sentence location and cue words including the phrases “in summary” and “in conclusion” from a document. These features are yet remained as the heuristic in the sentence scoring phase of many ATS system until now (Ferreira et al., 2014; Ferreira et al., 2013; Litvak & Last, 2013), and are referred in this study. Other important features including word-frequency, TF-IDF, sentence length and position, resemblance to the title and lexical similarity also have been extensively experimented by previous summarizer models such as in SUMBASIC (Nenkova & Vanderwende, 2005) and recent researches (Ferreira et al., 2013) where it has shown positive improvements in the content of the produced summary.

2.3.2 Machine Learning

Next, the machine learning technique can be divided into supervised, unsupervised and or semi-supervised approach. In a supervised approach, a trainable summarizer learns to select an important sentence from human summaries. For instance, a summarizer model introduced by Kupiec, Pedersen, and Chen (1995), generates a summary by classifying it into two classes of “summary” or “non-summary” sentence. Recently, a hybrid supervised summarizer model by Fattah (2014) has combined the maximum entropy model, naïve Bayes classifier, and a Support Vector Machine (SVM) approach to train and weight sentence features. Similarly, sentence features such as similarity of words and the importance of sentence-title overlap also have been identified by Ferreira et al. (2013). Their hybrid model was able to perform well using DUC 2002 dataset outperforming the lead-baseline model and shows that these sentence features are also language independent.

However, the supervised effort needs a lot of training and labelled data, hence differs from unsupervised techniques such as clustering that generate summaries based on discovered sentence cluster patterns or structure from the given document. For instance, one of the state-of-the-art summarizer named MEAD by Radev, D. et al. (2004) employed unsupervised centroid based approach, which identifies sentences that are highly relevant to an entire cluster of related documents. Meanwhile, (García-Hernández et al., 2008) extended the *K*-means clustering methods to find the best N-gram combination to represent important information from the text. On top of that, a co-trained summarizer by Wong et al. (2008) and a genetic algorithm approach by Litvak and Last (2013) have shown potential performance by optimizing the best linear combination of sentence features in developing their ATS model.

2.3.3 Semantic and Discourse-based

Latent semantic analysis (LSA) is a technique that finds relevant information in text where words and documents are mapped into a “concept” by observing the co-occurrence pattern of words. Here, semantically important sentences are identified for summary creation (Gong & Liu, 2001; Steinberger & Ježek, 2009). Nevertheless, the LSA approach still inherits the use of BOW language model where the order of words is not considered with high computation using SVD for a large set of data. On the other hand, deeper understanding in linguistics is needed for discourse-based approach in text summarization as applied in the study of Marcu (2000) using Rhetorical Structure Theory (RST) (Mann & Thompson, 1988). In RST, the relation between texts is mapped, which illustrates the sentence’s coherence relation. Marcu’s findings in discourse marker analysis have become the reference to other researchers. For example, the removal of certain phrases (discourse marker) such as “Furthermore” and “Moreover” has been widely used for English summarization.

2.3.4 Graph

Using graph approach, a vertex (nodes) can be used to represent text units such as words, phrase or sentences, and the edges link the related vertices. LexRank is a summarizer system developed by Erkan and Radev (2004) that connects two sentences if the similarity between them is above a predefined threshold. Meanwhile, a graph-based ranking algorithm TextRank by Mihalcea and Tarau (2004) works by using “vote” to cast one vertex to another where high votes indicate the importance of the vertex. A recent approach by Baralis, Cagliero, Mahoto, and Fiori (2013) exploited the use of Association Rules (AR) in Pattern Mining to discover the correlation between terms in a document using graph model. Their model showed improvement in comparison to heavy semantics-based models such as ontologies and deep NLP processing. On the other hand, recently Xie et al. (2017) try to improve the sentence scoring technique by merging the graph-based model with a new word-sentence relationship co-ranking model named CoRank. Their assumption is that each word should have a biased weight had shown superior results as compared to the baseline TextRank by Mihalcea and Tarau (2004) using Chinese news and DUC 2002 dataset.

2.3.5 Pattern-based

A pattern-based summarizer model tries to cater the issue of representing meaningful text unit from documents by using the discovered patterns without having to rely on prior or linguistics knowledge. For example using Frequent Pattern, Maximal Frequent Sequences and Closed Patterns representation. Previous researchers (Baralis, Cagliero, Fiori, & Jabeen, 2011; García-Hernández & Ledeneva, 2009; Ledeneva et al., 2008) and recent ones (Baralis et al., 2013; Qiang et al., 2016) have presented that a pattern-based model has the benefit to correlate the relationship between words by preserving the sentence semantically. The pattern-based natural representation produces encouraging results compared to the existing language model, which motivates the investigation of this study. Detailed discussion on the pattern-based model is catered in Section 2.5.

2.4 Topic Identification in Automatic Text Summarization

In order to identify the topic from a given text, an extractive summarizer model workflow mainly consists of three main tasks stated in (Nenkova & McKeown, 2011, 2012) that are consists of creating an intermediate representation of the input text, sentence scoring and sentence selection as illustrated in Figure 2.1, page 17.

2.4.1 Text Representation

After a text document has undergone the pre-processing task, a summarizer model will create an intermediate representation of the input content, which comprised the key source or topic of the text. One the most common approaches is using Vector Space Model (VSM) (Salton, Wong, & Yang, 1975), but other representations such as using a graph that was proposed by Erkan and Radev (2004) has also been exploited.

Representing a text as features involved two basic tasks, which are term indexing and term weighting (Lewis, 1992). In the term indexing task, the most representative term is assigned as the index of the document. Meanwhile, the term weighting task will assign an appropriate weight (usually Boolean, TF-IDF, term-frequencies and inverse document-frequencies) to the term index to measure the terms' importance throughout the document collection.

2.4.1.1 Term Indexing

There are variants of a language model that can be used as term index to represent a document(s) and sentence(s) as a feature vector in the VSM. The examples of terms index include the Bag-of-Words (BOW) (Kalogeratos & Likas, 2012; Le & Mikolov, 2014), the N-grams (Guthrie, Allison, Liu, Guthrie, & Wilks, 2006; Sidorov, Velasquez, Stamatatos, Gelbukh, & Chanona-Hernández, 2014; Tan, Wang, & Lee, 2002) and the pattern-based (Chim & Deng, 2008; Hernández-reyes, García-hernández, & Martínez-trinidad, 2006; Kim, Park, Lu, & Zhai, 2012; Li, Chung, & Holt, 2008; Ning, Yuefeng, & Sheng-Tang, 2012) model.

A. *Bag-of-Words*

A BOW representation is an individual word unit language model where documents are represented as a set of words contained along with the frequency. The general representation of a set of documents in D using BOW can be written as $D = (d_1, d_2, \dots d_n)$, where d_n is the document vector in the N number of document collection. The feature vector is the weight w_{dn} of each term index t_m denoted as $(t_1, w_{dn}; t_2, w_{dn}; \dots.; t_m, w_{dn})$ in document d_n.

The general problem in BOW is that, for example, a Malay news article regarding the Malaysia Airlines flight MH17 with the sentence "*MH17 ditembak oleh musuh*" and "*Musuh ditembak oleh MH17*", which brings about different meanings will have the same document representation in the VSM because of the same words being used, which are "*MH17*", "*ditembak*", "*musuh*" and "*oleh*". However, since the word order in BOW model is not preserved, it can lead to semantic issues and misleading meanings due to inaccurate representation (Kim et al., 2012; Le & Mikolov, 2014).

Nevertheless, the classic BOW approach, despite its semantic and word ordering issue, has seen much improvement. In solving the semantic issue, (Landauer, Foltz, & Laham, 1998) has introduced the Latent Semantic Indexing (LSI) based on the BOW model, which was applied in the area of Document Classification by Torkkola (2004) and Document Summarization by (Gong & Liu, 2001; Steinberger & Ježek, 2009). In LSI, words and documents are mapped into a "concept" by observing the co-occurrence pattern of words and related assuming words by their occurrences using the Singular Value Decomposition (SVD) technique. Even though LSI has the advantage on finding patterns from all documents without having prior knowledge, it