NEWS EVENT PREDICTION USING CAUSALITY APPROACH ON SOUTH CHINA SEA CONFLICT

TEO WEN LONG

UNIVERSITI TEKNOLOGI MALAYSIA

PSZ 19:16 (Pind. 1/13)

UNIVERSITI TEKNOLOGI MALAYSIA

DECLARATION OF THESIS / UNDERGRADUATE PROJECT REPORT AND COPYRIGHT

Author's full name :

Date of Birth :

Title :

Academic Session :

I declare that this thesis is classified as:

☐ **CONFIDENTIAL** (Contains confidential information under the Official Secret Act 1972)*

☐ **RESTRICTED** (Contains restricted information as specified by the organization where research was done)*

☐ **OPEN ACCESS** I agree that my thesis to be published as online open access (full text)

1. I acknowledged that Universiti Teknologi Malaysia reserves the right as follows:
2. The thesis is the property of Universiti Teknologi Malaysia
3. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
4. The Library has the right to make copies of the thesis for academic exchange.

Certified by:

SIGNATURE OF STUDENT	**SIGNATURE OF SUPERVISOR**
MATRIC NUMBER	**NAME OF SUPERVISOR**
Date:	Date:

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

"I hereby declare that I have read this final year project report and in my opinion this final year project report is sufficient in terms of scope and quality for the award of the degree of Bachelor Degree of Computer Science (Network and Security)"

Signature	:	_____
Name of Supervisor	:	Dr. Anazida Binti Zainal
Date	:	July 8, 2020

NEWS EVENT PREDICTION USING CAUSALITY APPROACH ON SOUTH CHINA SEA CONFLICT

TEO WEN LONG

A final year project report submitted in partial fulfilment of the
requirements for the award of the degree of
Bachelor Degree of Computer Science (Network and Security)

School of Computing
Faculty of Engineering
Universiti Teknologi Malaysia

JANUARY 2019

DECLARATION

I declare that this final year project report entitled *"News Event Prediction using Causality Approach on South China Sea Conflict"* is the result of my own research except as cited in the references. The final year project report has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature	:	
Name	:	TEO WEN LONG
Date	:	July 8, 2020

DEDICATION

This thesis is dedicated to my family, for their endless support in the success of my degree study. It is also dedicated to my beloved friends who taught and guide me throughout the project.

"Shoot for the moon, even if you missed, you'll land among the stars."

ACKNOWLEDGEMENT

First of all, I am grateful to my family for always backing me up during my degree study. I would like to express my warm appreciation and sincere thanks to them. No words can express my thankfulness to them.

I would also like to thank to my supervisor, **Dr. Anazida bt Zainal** for her patient and willing to teach. She is one of the best supervisor that always guide me whenever I faced with difficulties in the project. She also inspired me the beauty of data analysis and explore me with different techniques and platform that benefits to me throughout my study.

Besies, I would like to thank the authority of Universiti Teknologi Malaysia (UTM) for providing me with excellent equipment and study environments such as Students Lounge, High-speed WiFi, etc. Time spent in UTM is never wasted.

Last but not least, Thank you. To all the people I love.

ABSTRACT

South China Sea (SCS) is one of part from Pacific Ocean where generate huge economic value in fishing and shipping lane as well as high amount of natural resource. Various countries such as China, Vietnam, Philippines, Taiwan, Malaysia and Singapore surround SCS. Due to the strategy location of shipping lane and high revenue generated, SCS become place where several nearby countries compete for its territorial claims. Famous territorial disputes such as Spratly islands, Paracel island, Scarborough Shoal were happened for the claims of the wealth on SCS. To reflect these issues, newspaper are the main medium that propagate the first-line message to the public and update whenever there is SCS conflict happened. Within news events occurred, there are causal relation between cause and effect that able to obtain and analysis for the trend of event happening. This is known as causality in news article. Besides,in order to avoid any inevitable tragedy happen within SCS conflict, event prediction is important as it gives a better insight and foresee future events that might happen. Event prediction is a technique to measure the trend of happening events and forecast upcoming events that might happen. In this project, abstract event causality network proposed by Zhao *et al.* (2017) will be used as prediction model and furthermore embed the causality network into a continuous vector space.It is used because of its general, frequent and simple causality patterns as well as simplify event matching that suitable for event prediction. First, it extracts news article based on causality connector such as "because", "after", "lead to", etc into <cause, effect> tuple. Then, it represents the tuple with noun-verb representation and further generalised using WordNet and VerbNet. After that, an abstract causality network is build by using frequently co-occurring word pairs (FCOPA) and further embed into a continuous vector space for simplifying event manipulation while preserving cause-effect structure of the original network.

ABSTRAK

Laut China Selatan (SCS) adalah sebahagian daripada Lautan Pasifik yang menghasilkan nilai ekonomi yang besar dalam memancing dan lorong perkapalan serta jumlah sumber asli yang tinggi. Pelbagai negara seperti SCS, China, Vietnam, Filipina, Taiwan, Malaysia dan Singapura. Disebabkan lokasi strategi lorong perkapalan dan pendapatan tinggi yang dihasilkan, SCS menjadi tempat di mana beberapa negara berdekatan bersaing untuk tuntutan wilayahnya. Pertikaian wilayah yang terkenal seperti pulau Spratly, pulau Paracel, Scarborough Shoal telah berlaku untuk dakwaan kekayaan di SCS. Untuk menggambarkan isu-isu ini, akhbar adalah medium utama yang menyebarkan mesej baris pertama kepada orang ramai dan mengemaskini apabila terdapat konflik SCS. Di dalam kejadian berita berlaku, terdapat hubungan kausal antara sebab dan akibat yang dapat diperoleh dan analisis untuk trend kejadian yang berlaku. Ini dikenali sebagai kausalitas dalam artikel berita. Selain itu, untuk mengelakkan sebarang tragedi yang tidak dapat dielakkan berlaku dalam konflik SCS, ramalan peristiwa adalah penting kerana ia memberikan wawasan yang lebih baik dan meramalkan peristiwa masa depan yang mungkin berlaku. Ramalan acara adalah teknik untuk mengukur trend peristiwa yang berlaku dan meramalkan peristiwa akan datang yang mungkin berlaku. Dalam projek ini, rangkaian kausal sebab abstrak yang dicadangkan oleh cite zhao2017constructing akan digunakan sebagai model ramalan dan seterusnya membenamkan rangkaian kausal ke dalam ruang vektor yang berterusan. Ia digunakan kerana corak kausal yang umum, kerap dan sederhana serta memudahkan pemadanan acara yang sesuai untuk ramalan peristiwa. Pertama, ia mengeluarkan artikel berita berdasarkan penyebab kausal seperti "kerana", "selepas", "membawa kepada", dan lain-lain ke dalam <cause, effect> tuple. Kemudian, ia mewakili tuple dengan perwakilan kata nama-kata kerja dan lebih umum menggunakan WordNet dan VerbNet. Setelah itu, rangkaian kausal abstrak dibangunkan dengan menggunakan pasangan kata sepasang yang lazim (FCOPA) dan memasukkan lebih lanjut ke ruang vektor yang berterusan untuk memudahkan manipulasi peristiwa sambil memelihara struktur sebab-akibat rangkaian asal.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SCS	-	South China Sea
ORG	-	Organisation
LOC	-	Location
GPE	-	Geographical Entity

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Overview

Newspaper is an important part of our life as it is a printing media in which all information of either national or international news are published and delivered to the public every day. Henry Ward Beecher (1887), an American social reformer and well-known speaker once said, "Newspaper is a greater treasure to people than uncounted millions of gold." Articles within the newspaper play essential role in education development and makes public aware about events happening in the region or nation that they are living in. By reading newspaper, readers can learn and observe others point of view as it brings another whole new perspective on same events. In the digital era, artificial intelligence (AI) expert make use of the digital technologies to get a better insight within the traditional media, newspaper.Baker (2019) Within these digital technologies, event prediction held a big portion as it forecast future events and it is valuable to alert public on predicted events.

Event prediction is a data analytic technique that make use of experience and knowledge as well as pattern from past to predict future events. Natural disaster prediction is one of the examples in applying concept of event prediction. For example, Japan is a well-known earthquake active country as its archipelago is in an area where multiple continental and oceanic plates collapsed together. Hence, earthquake forecasting is important for Japan and scientific report (D.Goltz, 2018) stated that earthquakes cluster in time and location as it can be predicted and take precaution before tragedy happened. Besides, event prediction is also useful in business intelligence. A report in 2016 also showed that business analysis and prediction help them to understand more about their customer, in order to enhance the success of their marketing strategies. (Erevelles *et al.*, 2016)

Every country had its obligation to protect its national security. National security is a requirement to maintain the survival of a country though economic, diplomacy, political and ethical power and focus on freedom from military threat and political coercion. Without national security, a country might at risk and attacks such as terrorism, sabotage, information warfare, etc might infiltrate the country. For example, ISIS threat stunned the world as a gunman, Mehdi Nemmouche opened fire at Jewish Museum of Belgium in Brussels as he is suspected in joining extremist groups, ISIS in Syria. This event took 4 lives of innocents.(Casert, 2016)

South China Sea (SCS) is a conflict zone whereby an estimated USD5 trillion worth of raw products shipped through shipping lanes in SCS each year and its nearby countries made them fight over each other to have the main control of the whole SCS. (Fensom, 2016) The conflict is known as South China Sea disputes. The events of territorial disputes of South China Sea populated all the newspapers and many events regarding to the disputes were reported through national news agency. It rises concerns about the beginning of world war as for example a near-collision between US warship, Decatur and Chinese Luoyang missile destroyer in South China Sea highlights the escalating danger of confrontation between US and China. (Ni, 2018) SCS is strategically located at peripheral ocean that is a piece of the Pacific Ocean, starting from Karimata and Malacca Straits to the Strait of Taiwan with area about $3\,500\,000\,km^2$.Besides, South China Sea is rich in marine life and natural resources such as oil and natural gas, even have the most of the important shipping lanes in the world. (E.Hayes, 1980)

Causality is the relationship between cause and effects. Every event will occur first on cause and followed by effects. In SCS disputes, causality is highlighted between benefits from SCS (cause) and territorial disputes (cause) is clearly highlighted in SCS disputes events.

In order to have an advanced insight among these disputes, event prediction is necessary, and causality should be taken as main attributes. However, there are still several problems and challenge to be solve in order to achieve an excellent prediction model based on causality.

1.2 Problem Background

National security is always the top priority of governments to protect society from disruption owing to a disaster or crisis. There are many aspects on national security such as territorial, economic, physical, social, political etc. However, the peaceful of national security had been affronted by South China Sea disputes.

Due to geological and resources advantages of South China Sea, countries within the region such as Brunei, China, Taiwan, Malaysia, Indonesia, Philippines, Vietnam etc. made competing the territorial claims over it. Based on news on The National Interest in 2016, an estimated US 5 trillion worth of global trade passes through the South China Sea annually. Hence, territorial disputes in the South China Sea started to concern worldwide community about peace of world. In order to claim the ownership of South China Sea, countries are challenging against each other by putting military force in the area. This can be observed from the news of China spent almost 1 year to build 7 new islands by moving sediment from the seafloor to reefs and after that focused on building ports, airstrips and other military structures on the islands.

South China Sea dispute had a brief background involving timeline from 221 BC until recent. Each of the historical event occurs and accumulates and eventually things go haywire. Many dispute events happened in either small or large scale. For example, Spratly island dispute (Gonzales, 2014) and “nine-dash” line (Zhen, 2014) that proposed by China is some of the significant disputes in South China Sea. Besides than these two issues, there are many issues that remain unsolved and will constantly concerning the worldwide community.

However, all the information retrieved from news article are unstructured. Unstructured data have no recognizable structure via pre-defined data models and schema and mainly generated by human or machine. (Taylor, 2018). By collecting these unstructured data from the past and analysing its trends, we are able to have a better understanding about what may happen in the future. (Bell, 2016). In SCS disputes, event prediction is important to give public a better understanding about future events that might happen. A better policy can be made with regards to protect

national security under SCS disputes with the event prediction technique based on unstructured data in news articles.

In event prediction based on news articles, there may have some challenges. First, news article is a type of unstructured data that contain a lot of valuable information in term of cultural, social and historical (Yzaguirre *et al.*, 2016) but doesn't fit into traditional row and column structure of relational databases.(mongoDB, 2016). It require substantial manual effort to analyse and extract the essential information from news articles. Second, a event that causes another events may completely different from the real prediction. It is indicating that the predictive model provides a faulty outcome that hard to distinguish from the true prediction.

There are several researchers research on topic event prediction with different method. Granroth-Wilding (2016) proposed a predictive neural network model that learns embeddings for words describing events, a function to change embeddings into event representation and a function to predict the degree of relationship between two events.However, the model is more focus on chain or events sequence which is good for rich-infomrative events but might not suitable for news articles that have unordered sequence. Preethi (2015) proposed an event prediction model for Tweets using temporal sentiment analysis and causal rules extraction. This model is useful to analyse user's sentiments and predict future events using temporal attribute. This study analyse sentiments of user's opinion and is not suitable for news articles whereby formal news reports seldom express their sentiments within the articles.

The current research done is more focus on causal event detection and extraction, which related to effective distributed word or sentence representation. Mikolov *et al.* (2013b) had proposed a state-of-art framework, Word2Vec for distributed word representation. However, Word2Vec is limited to those words that are morphologically similar where Word2Vec embed every word as an indepenedent vector.

Thus, this research is aimed to provides a prediction framework that can actually address these problem by using state-of-art sentence representation and

sentence similarity. In this research, we train and compare 3 different sentence representation technique (Doc2Vec, InferSent, and BERT) as well as their sentence similarity with the input queries in order to get the most possible predicted output.

1.3 Problem Statement

South China Sea (SCS) is a conflict zone where events happened from time to time with different severity. This greatly impact or influence the policy that made by government of Malaysia to overcome the negative impacts brought by SCS territorial disputes in term of national security. The problem is to extract valuable information from SCS conflict events and predict the future events that may happen. Besides, online news is unstructured data and extracting correct information from massive online news articles that contain different resources automatically is part of the challenges. Event prediction is biased to measure, and a good predictor is needed to make sure that its prediction is accurate and precise.

1.4 Aim/Purpose of Study

This study will address the SCS disputes issues by comparing sentence similarity based on different sentence representation technique such as Doc2Vec, InferSent and BERT. These technique helps to predict the most possible event that follows the nature of event causality and provides simplify event matching which is advantage for event prediction.

1.5 Research Question

1. How to obtain useful information from news article that related to South China Sea conflict?
2. Which sentence representation algorithm is suitable to represent causal event?
3. How to use the extracted information to train and build a prediction model?

1.6 Objective

1. To extract <cause, effect> causality pairs from news articles in South China Sea conflict by using causality connectors
2. To compare sentence representation algorithm by comparing sentence similarity between input queries and pre-defined <cause, effect> causality pairs.
3. Build a prediction model based on highest sentence similarity score for possible causality output event.

1.7 Scope

Event prediction is important to have better understanding for future event that may happen, but it also have its limitations that need to address and narrow down. Below are the scope of this study:-

1. The coverage of this study will be using news articles that extracted related to SCS conflict only.
2. The study will focus on SCS conflict news from online news article that having keywords related to SCS.
3. The study will only covers the some of the causality connectors such as "because of", "because", "after", "due to", etc.
4. The study will only covers 3 sentence representation algorithm which are Doc2Vec, InferSent, and BERT.
5. Prediction model in this study will be based on sentence similarity for the possible causality output event.

1.8 Significant of Study

Event prediction is important because it give a better insight and forecast for public about events that possible to happen. Besides, in SCS conflict, event prediction can be used to protect national security of Malaysia and appropriate actions can be took by governments to prevent the happening of unseen tragedy based on predicted events.

In addition, Malaysia governments are able to protect the sovereignty of country in SCS region to claim peace and avoid involving in SCS conflict.

1.9 Organisation of Study

Chapter 2 will be discussed the literature review of event prediction and information extraction from news articles based on different method and referenced method. **Chapter 3** will focus on research methodology, research framework, overall research phase, as well as measurement and rules. **Chapter 4** will be present about initial result of implementing prediction model proposed by Zhao *et al.* (2017) by on step-by-step approach with causality attributes. **Chapter 5** will be summarised work done in PSM 1 and discussed future works involving in this project.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter will discuss the literature review of event prediction using causality approach on South China Sea (SCS) conflict. First, the chapter will briefly explain about event prediction and SCS conflict Then,the chapter will discuss text mining which consists of many stages such as text preprocessing, event relation, event detection, event extraction, as well as event extraction. After that, this chapter will briefly discuss the prediction model technique and justify the most suitable algorithm for SCS conflict event prediction in this paper. Finally, this chapter will compare existing work that had done in the domain of event prediction. Last but not least, this chapter will discuss open issues in event prediction that need to be addressed, followed by a short brief summary.

2.2 Event Prediction

Event prediction is a technique to measure the trend of happening events and forecast upcoming events that might happen. According to Cambridge dictionary, event stands to anything that happens especially something important or unusual while prediction is a statement about what you think will happen in the future. Figure 2.1 shows a taxonomy of event prediction. The taxonomy is sketched based on generic steps needed to be performed in event prediction. Steps such as text preprocessing, event extraction, event representation etc. is important in the progress of building up a successful predictive model. The main part of event prediction is text mining which will be further discussed in the following section.

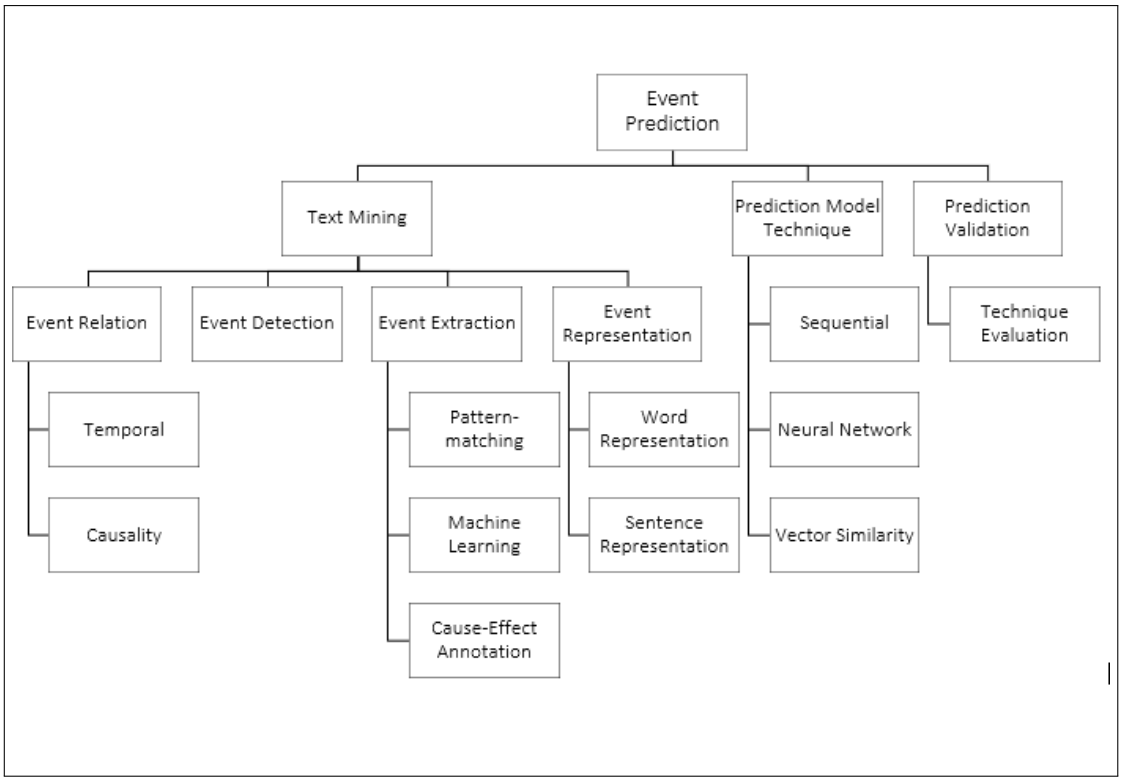


Figure 2.1 Taxonomy of Event Prediction - Generic steps needed to performed in event prediction

Event prediction involves in many domains. Research on areas of event prediction are active and wide. Example of domains in event prediction are below:

1. Natural disaster such as earthquake (Asencio-Cortés *et al.*, 2017; Asim *et al.*, 2017) and tsunami (Mulia *et al.*, 2016),
2. Political such as election prediction (Tung *et al.*, 2016)
3. Social such as protest event in Argentina (Ning *et al.*, 2016)
4. Economics such as predicting market stock price (Ding *et al.*, 2015; Vargas *et al.*, 2017), etc.

In news articles, there are many events happen everyday. From political issues, to social problems, economic trend and entertainments, newspaper provides public daily updates on these events. Recently, conflict happened on South China Sea (SCS) has becoming popular and continuously escalated. The happening of SCS conflict

concerns countries nearby the regions and it must be solved and actions should be taken to prevent happening of any tragedy.

2.3 South China Sea Conflict

South China Sea (SCS) is located strategically within Asia countries such as China, Taiwan, Indonesia, Philippines, Vietnam, and Malaysia. It has the busiest shipping lane as one-third of world's shipping passes through SCS and almost 3.37 trillion global trade happened within SCS in year 2016 (ChinaPower, 2016). Due to these huge market profit, many countries started to claim that SCS is their own territory and China even illustrated a "nine-dash line" which is a huge part from SCS and claim that region within "nine-dash line" is China's territory (Zhen, 2014). These started the dispute between countries. For every dispute, public are able to know the flow of events through newspaper.

Examples of famous SCS dispute are Spratly islands dispute and China-Vietnam Military Clash. Spratly islands covered with huge amount of natural resources, global maritime areas and commercial shipping lane. In year 2015, China had militarised Fiery Cross Reef, reef located at the western edge of SCS and centre of Spratly islands by constructing military-level airstrip and seaport. This made concerns to Vietnam and Philippines as they felt threaten to their sovereignty in SCS region. Besides, China and Vietnam are continuously declared their ownership in SCS on oil and gas exploration issues. The issues are escalating as reports (news, 2014) showed that Vietnam tried to ram Chinese vessels in SCS dispute area.

New articles are written by different news agency with different perspective and opinion.Among many of the newspaper agency, National news agency such as Xinhua News Agency for China and Vietnam New Agency (VNA) are said to be the voice of government. Battistella (2005) stated that Xinhua is the biggest propaganda machine that spread the core concept of China's government to the public. Hence, there are huge amounts of valuable information that represent and reflect the action of government can be obtained within news article from national news agency. However, without proper technique to extract useful information from the news article, the data source will be difficult to be obtained and time-wasting. In the next section, text mining

will be introduced as an effective method to obtain valuable information from the news articles.

2.4 Text Mining

Text mining, also known as knowledge discovery from textual databases and related with steps of obtaining valuable and non-trivial pattern or knowledge from unstructured text documents (Tan *et al.*, 1999). A general framework proposed by Tan *et al.* (1999) in Figure 2.2 shows text mining contain two major elements, which are text refining and knowledge distillation. Text refining are techniques to transforms free-form text into intermediate form meanwhile knowledge distillation is to obtain the pattern or knowledges within the intermediate form. From text refining, two types of intermediate form can be obtained, document-based or concept-based which are able to be further processing into visualisation or predictive modelling.

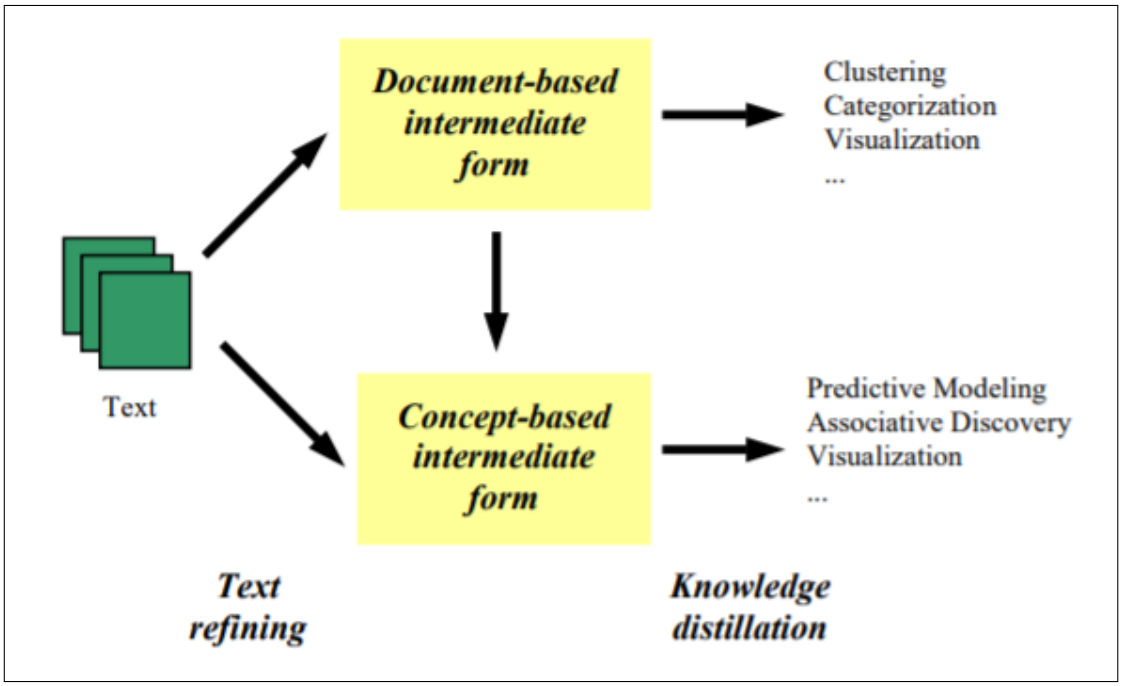


Figure 2.2 General Text Mining Framework proposed by Tan *et al.* (1999)

However, there are open problems in text mining that need be addressed. For examples, variety of intermediate form that causes complexity and uncertainty in text mining, multilingual text refining, domain knowledge integration (Lima *et al.*, 2009), and customized autonomous mining (Afzal *et al.*, 2010).

2.4.1 Types of Data

With text mining, useful information can be extracted for further processing. However, the volume of data is dramatically increasing nowadays, associated with high flow of data and high variety of information. For example, inside a single webpage, different type of data such as images, audio, textual data, comments, etc. can be obtained and categorised. Two major categories of data are structured data and unstructured data. Both of the categories will be further discussed in the following sections.

2.4.1.1 Structured Data

Structured data is referred to the data that organisation all the information in formatted way and easily to extract and analysis by relational databases(Taylor, 2018). Example of structured data are phone number, name, identification numbers, date,etc. Format of structured data is analysable with human-generated queries and algorithm.

2.4.1.2 Unstructured Data

Unstructured data is referred to the information that does not have a pre-defined model or structure and unable to store in traditional relational databases. Example of unstructured data such as text, emails, blogs, web pages, images, audio, comments on social media, etc. has no proper structure and has to be processing into valuable information before analysing them. Figure 2.3 shows the example of structure data(numerical and statistical data) and unstructured data(text, audio, video and blog posts).

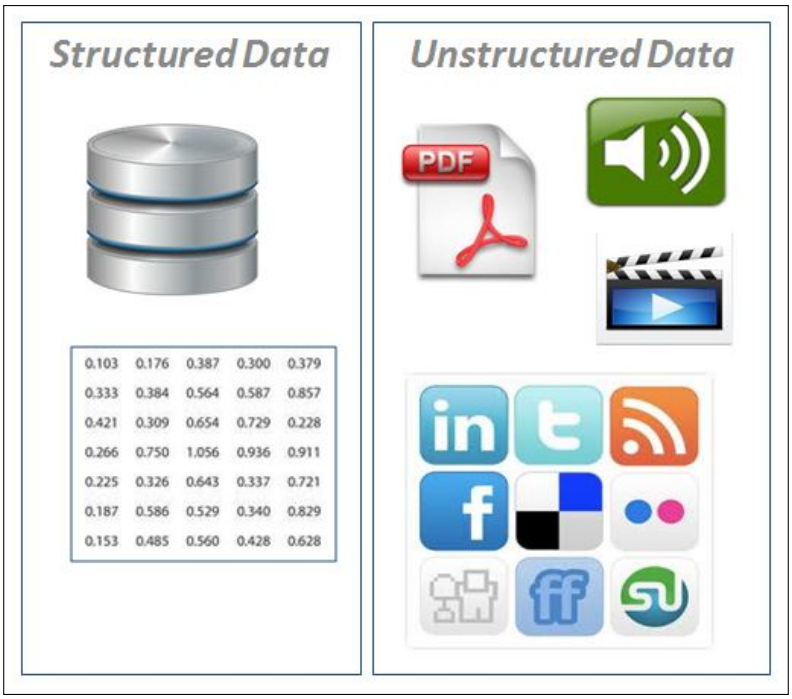


Figure 2.3 Example of structured and unstructured data (Michael Benz, 2017)

Nowadays, unstructured data is getting enormous. Information contained within these unstructured data originating from humans is absolutely richer and more valuable than traditional numeric data sources (Frunza, 2016). Besides, study in 2019 (Kulkarni, 2019) showed that 95 percent of businesses need to manage unstructured data and more than 150 zettabytes(150 trillion gigabytes) of data will be needed for analysis by 2025. Because of that, exploring and extracting useful information from these unstructured data is essential. However, text mining is difficult due to ambiguity of languages, multiple duplicated words with different meanings, abbreviation, linked references etc(Bhardwaj, 2016).To overcome the difficulties, many techniques such as Named Entity Recognition (NER), Nature Language Processing (NLP), sentiment analysis, causality relationship extraction etc. are developed to extract high quality information from textual data.

News event from SCS conflict is one of the unstructured data which the textual data is raw and does not have a proper structure on the sentences. Hence, further processing for the news events is important and it should be started with text preprocessing.

2.4.2 Event Relation

Event relation is important in text mining as it provides better insight and understanding about event happens either before or after. Generally, there are 2 major type of event relation that can be observed and obtain from news article which are temporal relation and causality relation. As events happen from time-to-time, temporal attribute should be focused as it measures the trend and evolution of event. Besides, there is always a cause for events to happen, hence causality is used to relate events from the very beginning based on its cause.

2.4.2.1 Temporal Relation

Temporal is indicated as time-based or using time as the main measurement. Usually, events getting worse from a small case. For example, in medical domain, long-term follow-up is important for medical investigation. Maziarz *et al.* (2017) make use of temporal relation to observe the trend of patient's health condition over time, and eventually predict patients' risks for a future adverse outcome. Besides, temporal relation had been widely adopted in event prediction with different domain such as micro-blog and tweets (Preethi *et al.*, 2015), sports (Grolinger *et al.*, 2016), stock market flow (Ding *et al.*, 2015) and etc.

2.4.2.2 Causality Relation

The main concept of causality is that one or more things/events as causes can cause one or more things/events to happen as effect. In news article, there are many events that related with each other but hard to be observe and detect by human or massive news events happen in the same time cause human hard to digest all the news happened in one time. Thus, people tend to automated the extraction of news article based on causality. However, there are many challenges in automating the process. For example, Pechsiri and Piriyakul (2010) faced difficulties on explanation knowledge graph through causality extraction from text such as causal-boundary determination and effect-event pattern determination. Nevertheless, it is useful for event extraction because it contain cause and effect that allows people to understand more about the following events. In this study, we will focus on causality relation as SCS conflict happens after certain small case instead of following time flow.

1. **Causality Pair**

   Causality pair is also known as entity pair, used to represent causality within pairs. For instance, Mirza (2014) stated possible causality pairs in his study as (1) main events of consecutive sentences, (2) pairs of events in the same sentence, (3) an event and a time expression in the same sentence. If the pairs are tagged as (e_1,e_2), the pairs will be *events-events*, *events-timex*, *timex-timex*.

2. **Causality Rules**

   After identify the potential cause-effect pairs, it is difficult to execute human-annotation on these causality pairs as the amount of data is getting enormous. Hence, autonomous annotation is proposed. In the work of Radinsky *et al.* (2012), Radinsky use causality rule to predict the future event as abstraction tree (AT) had been build earlier and using causality rules, a causality graph is built to explain the predicted event clearer.

   Moreover, in order to get an accurate categorised result based on causality pair, causality rules is set in every text extraction. Zhao *et al.* (2017) construct a set of causality rules that follows the template of <**Pattern,Constraint,Priority**> where Pattern indicates the trend of events between causality pairs, Constraint indicates the limitation on sentence which the pattern can be applied, Priority indicates the priority of causality pairs to be execute first or second. After extracting causality pairs from news article as well as generalise for more general purpose, an abstract causality network is built.

2.4.3 Event Detection

Event detection is a techniques to discover occurrences of events by analysing text stream in the sources. Generally, event detection can be broadly categorised as either Feature-pivot (FP) or Document-pivot (DP) (Fedoryszak *et al.*, 2019). Based on Chen *et al.* (2019), feature-pivot is the method that detecting abnormal patterns in appearance of features such as words. When an event happens, the frequency of a feature will be abnormal comparing to its unusual behaviour which indicating a potential new event. Meanwhile, document-pivot is using documents as items to clustered them by using similarity measure.

The main idea of feature-pivot event detection is to represent the events as number of features that showing an abnormality in appearance counts.It is illustrated in Figure 2.4. These method rely on large amounts of dataset as the unusual frequency will be taken as important features.This technique is widely used in many domains especially in news documents. For example, Mele *et al.* (2019) use this technique to first extracts event features in tweets as well as clusters the documents according to the features. By finding a frequent pattern based on an infinite-state automation, it helps to model the changes in word frequency, state transitions as well as the events. This technique will also used in this project as Causality Mention (CM). By referring causality connector such as "because', "because of", "lead to" and "after" from Zhao *et al.* (2017), the sentences that tagged with these connectors are selected and initially identified as event.

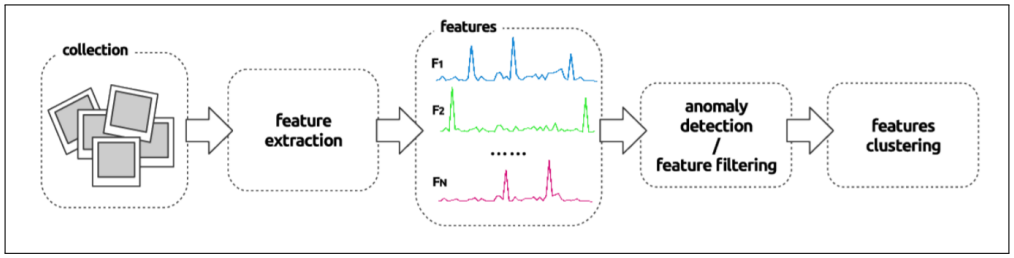


Figure 2.4 Feature-pivot paradigm for event detection, adopted from Schinas *et al.* (2018)

Document-pivot is a method where cluster documents based on their semantic similarity and group them into events. It is illustrated in Figure 2.5 This method is a bit different from traditional Bag-of-Words approach as it helps to identify event-related data that mixed with noisy data apart from clean data. TwitterStand (Sankaranarayanan *et al.*, 2009) is a news processing system using document-pivot approach to capture tweets corresponding to breaking news. However, this approach as low performance due to fragmentation of news topics, noisy content and the limited power of Bag-of-Words in short texts as tweets.

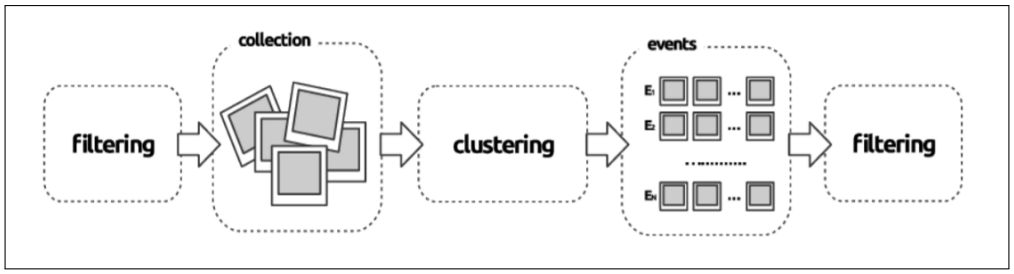


Figure 2.5 Document-pivot event detection, adopted from Schinas *et al.* (2018)

2.4.4 Event Extraction

Event extraction is one of the common application of text mining as it obtains specific knowledge concerning incidents referred to texts (Hogenboom *et al.*, 2011). Figure 2.6 shows the general procedure of event extraction. Event extraction uses data that are preprocessed using the methods mentioned in the previous chapter, and further represent by general causality pairs.

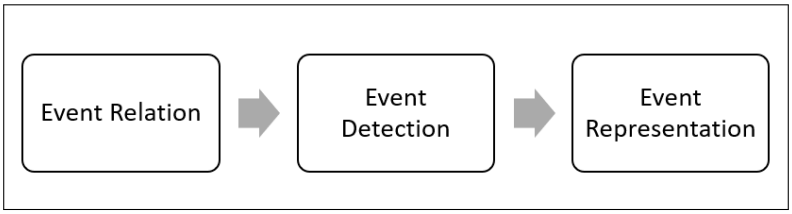


Figure 2.6 General procedure of event extraction

Based on Xiang and Wang (2019), event extraction is defined as one of the most important and active research domain in Naturral Language Processing. The author break down this big domain with several representative methods such as Pattern-matching, Machine Learning and also Annotation based on Cause and Effect. The following subsection will further discuss the concept and techniques' solution approaches.

2.4.4.1 Pattern-Matching

Pattern-matching is a traditional event extraction method and require a specific template or structure to extract important information from the corpus. In order to obtain important insight from unstructured text, linguistic pattern is used to automatically build up a event patterns within the corpus. Figure 2.7 shows 13 linguistic patterns and example that applied to extract important information from corpus. For example, if given the a sentence, "victim was murdered". Based on the

first linguistic pattern, "victim" will be tagged as <subject> while "murdered" will be tagged as verb.

Figure 2.8 illustrate the process of event extraction based on pattern-matching. By following the linguistic pattern template, the sentence "They took 2-year-old Gilberto Molasco, son of Patricio Rodrigues" will be extracted with "took" as the trigger of the whole sentence and "Gilberto Molasco" as the subject or victim. Pattern-matching is domain-specific and works well in specific dataset. However, it does not provide accurate extraction in unstructured data such as news article and post in social media.

SN	Linguistic Pattern	Example
1	<subject> passive-verb	<victim> was murdered
2	<subject> active-verb	<perpetrator> was bombed
3	<subject> verb infinitive	<perpetrator> attempted to kill
4	<subject> auxiliary noun	<victim> was victim
5	passive-verb <dobj>	killed <victim>
6	active-verb <dobj>	bombed <target>
7	infinitive <dobj>	to kill <victim>
8	verb infinitive <dobj>	threatened to attack <target>
9	gerund <dobj>	killing <victim>
10	noun auxiliary <dobj>	fatality was <victim>
11	noun prep <np>	bomb against <target>
12	active-verb <np>	killed with <instrument>
13	passive-verb <np>	was aimed at <target>

Figure 2.7 Example of Linguistic pattern and its pattern, adopted from Riloff *et al.* (1993)

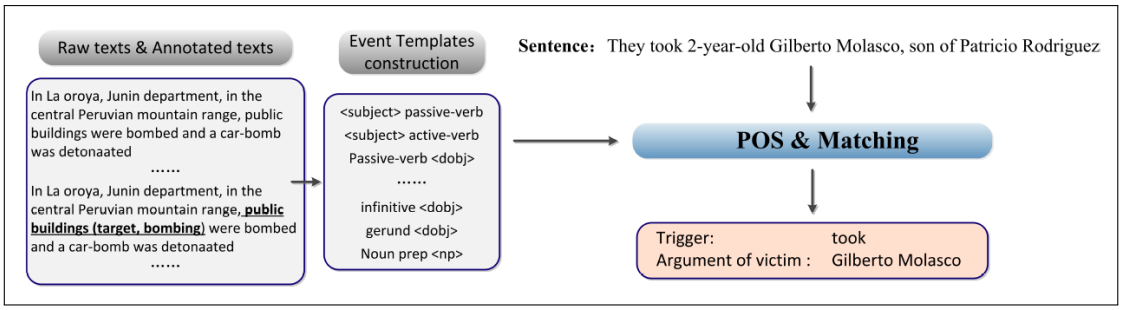


Figure 2.8 Example of event pattern construction and event extraction based on pattern matching adopted from Xiang and Wang (2019)

2.4.4.2 Machine Learning

The important difference between pattern-matching based and machine learning based event extraction is that machine learning based event extraction rely on algorithm such as Support vector Machine (SVM), K-means clustering, etc. for the extraction. Apart of defining lingustic pattern, machine learning based event extraction learns classifier from training data based on feature engineering work. Feature engineering is a technique about creating new input features based on the existing one. In natural language processing, most of the comoon features can be divided as 3 types, *lexical*, *syntactic* and *semantic* features.

Lexical feature usually includes word lowercase and stopword removal, text lemmatization, Part-of-Speech (POS) tagging while syntactic features are extracted based on dependency parsing. The general idea of dependency parsing is to create edges between words in a sentences that denoting different types of relations. For example, Universal Dependencies (de Marneffe *et al.*, ????) documented standard dependencies for cross-linguistic typology such as nsubj, nmod, obj, etc. Figure 2.9 shows the example of dependency parsing on simple sentence, "Ivan is the best dancer". From this sentence, "dancer" is the root word and "Ivan" is the nominal subject (nsubj) to "dancer". Nominal subject is a noun phrase which is the syntactic subject of a clause. Besides, copula (cop) is the relation between verb and verb, determiner (det) is the relation between head of noun phrase and determiner and adjectival modifier (amod) is a adjectival phrase that represent the meaning of noun phrase. The representation helps to describe the grammatical relationship in a sentence so that researcher can understand more on the sentence. Lastly, semantic feature is normally includes obtaining synonyms of words as well as argument identification.

Figure 2.10 illustrate a overview of event extraction based on machine learning approach. From the figure, each of the features are embed into vector which is also another important concept that will be discussed on the following subsection.

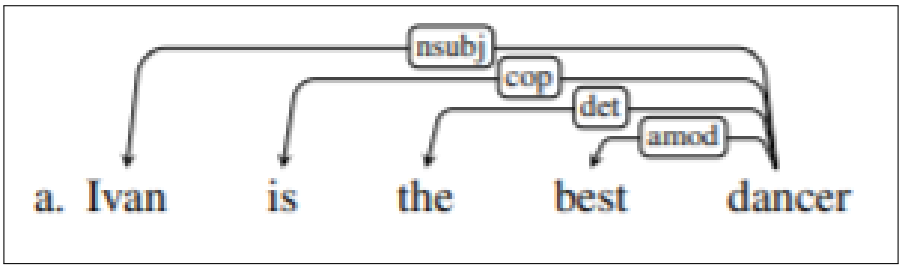


Figure 2.9 Example of Dependency Parsing Visualisation

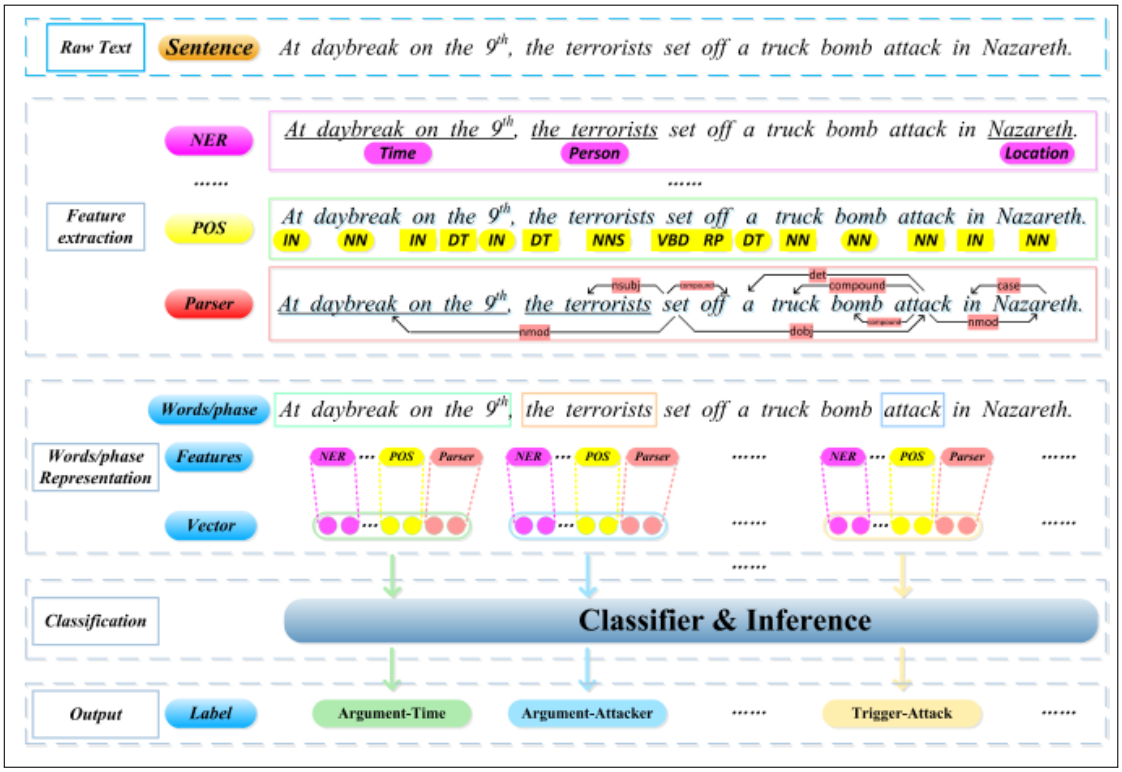


Figure 2.10 Example of event extraction based on machine learning

2.4.4.3 Cause-Effect Annotation

Annotation is an important concept in obtaining useful information in Natural Language Processing (NLP). By annotating raw text data, all the text become meaningful as tag with specific characteristics. For example, a Named Entity Recognition (NER) annotation can simply help researchers to eliminate most of the noise in the raw data by only obtaining Person, Location, Time, etc. The most widely used annotation tools for NLP is Brat and WebAnno. Figure 2.11 shows the screenshot

of Brat annotation tools. By using Brat, cause and effect can be easily annotated followed by their relationship (cause-effect).

One of the well-known test data for cause-effect annotation is SemEval-2010 Task 8 Hendrickx *et al.* (2010). It is a standard test data for multi-way classification of semantic relations between pairs of nominals. In this dataset, there are many semantic relations annotated such as Cause-Effect (CE), Instrument-Agency (IA), Product-Producer (PP), etc. In this research, CE will be focused and the extracted cause and effect will be taken as test data for model validation.

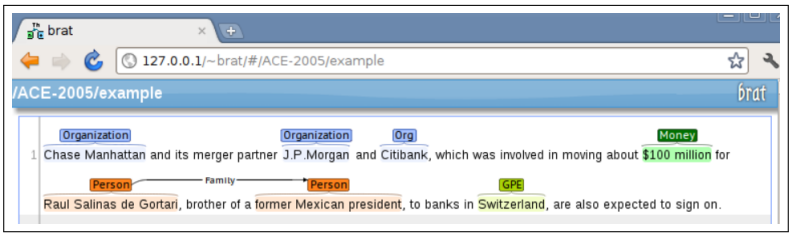


Figure 2.11 Screenshot of Brat Annotation Tool

Table 2.1 shows the recent research for event extraction as well as their approach and limitation. Most of the research is based on machine learning approach since it provides a feed forward neural network for continuous word or sentence representation. It might be helpful in dealing with large raw text corpus and highly unstructured data.

Table 2.1 Comparison of Recent Research on Event Extraction

Author, year	Domain	Approach	Limitation
(Valenzuela-Escárcega *et al.*, 2015)	Domain-independent Rule-based Framework for Event Extraction	Pattern-Matching	Needs to design multiple rules for different domain
(Ritter *et al.*, 2012)	Open domain event extraction from twitter	Machine Learning	Segmentation error on unknown words on Tweet
(Wu *et al.*, 2017)	Event Timeline Extraction on News Corpus	Machine Learning	Require large text corpus
(Vossen *et al.*, 2016)	Using knowledge resources in cross-lingual reading machine from news	Machine Learning	Require knowledge acquisition and NLP improvement on a massive scale
(Rospocher *et al.*, 2016)	Building event-centric knowledge graphs from news	Annotation	Expert knowledge required to build knowledge graph

2.4.5 Event Representation

Event from news article usually construct based on sentences or phrases. However, unstructured raw text is highly ambiguous and without proper rules or structure, machine wouldn't know any meaning from these text. (Mikolov *et al.*, 2013b) proposed a method called distributed representation of words in vector space or word embedding that actually helps natural language processing to achieve better performance in term of linguistic understanding. These word embedding approach is useful in term of capturing semantic and syntactic patterns between words. Below are some example of state-of-art word embedding technique:

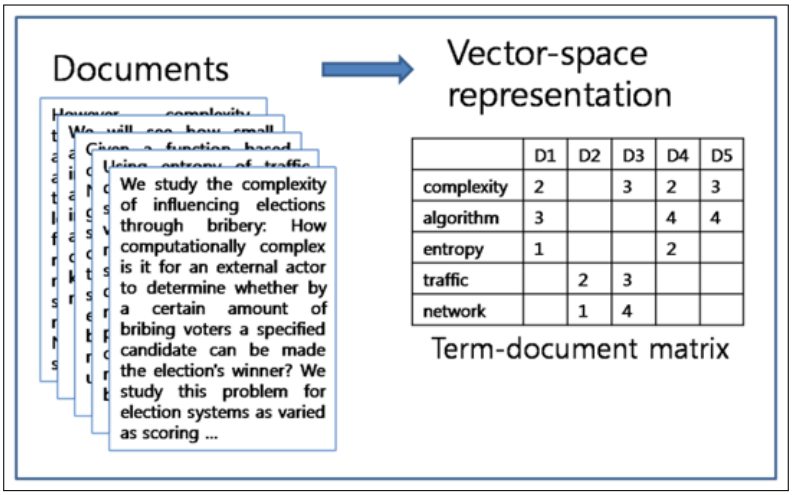
2.4.5.1 Word Embedding

In word embedding, the most commonly used technique is Latent Semantic Analysis (LSA), Word2Vec and Glohal Vector (GloVe).

1. **Latent Semantic Analysis (LSA)**

   LSA is a statistical technique by using bag-of-words where it convert every words in document into vector by counting the number of occurrences of each words in the documents. Then, a term-document matrix (TDM) is constructed

for the document vectors. The core concept of LSA is term frequency where it take weighting scheme (boolean of O-term not used and 1-term is used) into consideration. Figure 2.12 shows an illustration of TDM and a singular value decomposition is used to get the most closest vector embedding in the subspace.



	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Figure 2.12 Term-document Matrix

2. **Word2Vec**

Word2Vec is first proposed by Mikolov *et al.* (2013b) as a solution for neural network based training of word embedding. Until now, Word2Vec is becoming the state-of-art word embedding algorithm as many of the research extends from this concept such as Zhang *et al.* (2015), Lilleberg *et al.* (2015), etc.

There are 2 different approaches in Word2Vec, Continuous Bag-of-Words (CBOW) and Skip-gram model. Figure 2.13 shows illustration of CBOW and Skip-gram in Word2Vec. CBOW is trained and learning to predict the word by context as well as maximize the probability of the target word by finding the context. For example, given a context "yesterday was a really [...] day". CBOW model will be about to predict the possible target word as "beautiful" or "nice". On the other hand, skip-gram model is a reverted version of CBOW as it is used to predict the context. Given a word "beautiful", there are higher possibilities that "yesterday was really [...] day" will be selected as the context.

The most common downstream application that applied Word2Vec algorithm is smartphone keyboard. Word2Vec helps to implement next-word prediction

feature to ease smartphone user during their text typing. Besides, Word2Vec also helps in getting valuable information from customer reviews. Business can use this technique to analyze survey responses and perform business analytic on these data.

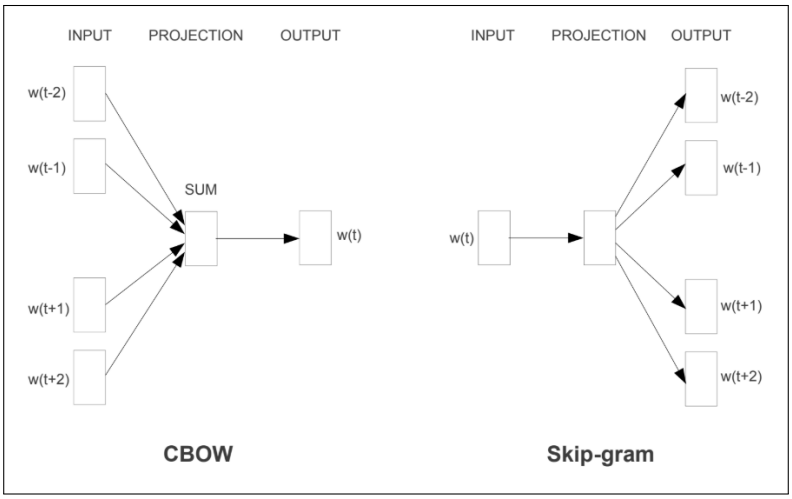


Figure 2.13 Word2Vec training model, adopted from Mikolov *et al.* (2013a)

3. **Global Vector (GloVe)**

GloVe stands for Global Vector for Word Representation. It was proposed by Pennington *et al.* (2014) and soon become one of the state-of-art word embedding model due to its faster training speed by using non-zero matrix entries. GloVe is training on word co-occurrence matrix and its probability and ratio between target words and context words from 6 billion token corpus. Besides, GloVe provides a pre-trained word vectors for different domains such as Wikipedia, Common Crawl and also Twitter. All of the pre-trained model will be trained in different vector dimension to fix the usage of researchers. In this research, GloVe pre-trained word vectors will be one of the reliable resource in loading word vectors since training word vector from scratch is time-consuming and relatively expansive in computational resource.

However, word embedding seems not sufficient when it comes to phrase or sentence level. For phrase or sentence level embedding, most of the approach will be simply averaging a sentence's word vector/ Bag-of-words approach. This brings to

several constraint in word embedding such as semantic similarity between sentences. word embedding only consider to represent the meaning of 1 word into vector and when comes to phrase and sentence, it cannot represent accurately the meaning of phrase or sentence. To address this problem, sentence embedding technique is used. The following subsection will discuss several sentence embedding technique such as Doc2Vec, InferSent and BERT.

2.4.5.2 Sentence Embedding

1. **Doc2Vec**

   Doc2Vec is proposed by (Le and Mikolov, 2014). It is also known as Paragraph Vector as the algorithm added a Paragraph ID on its training. Based on Word2Vec approach, Doc2Vec follows the concept of CBOW and skip-gram, additionally added another vector (Paragraph ID). There are also 2 method that similar to CBOW and skip-gram in Word2Vec which are PV-DM (Paragraph Vector - Distributed Memory) and PV-DBOW (Paragraph Vector - Bag-of-Words). Figure 2.14 shows the illustration of PV-DM and PV-DBOW model. PV-DM predict the missing target word in the context while PV-DBOW predict all the context from the target word.

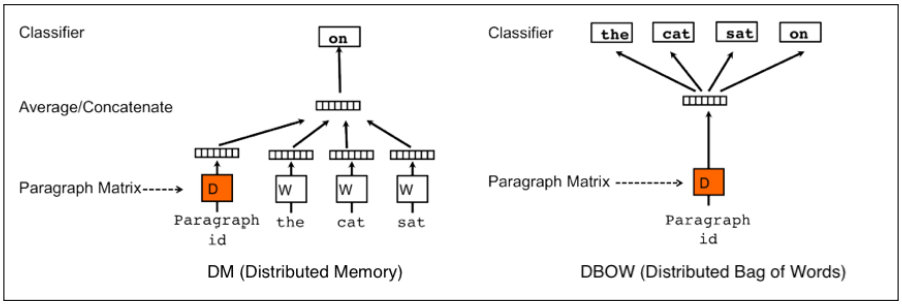


Figure 2.14 PV-DM and PV-DBOW model in Doc2Vec (Le and Mikolov, 2014)

Doc2Vec is a unsupervised learning algorithm that learns vector representation for variable-length pieces of texts. Doc2Vec had achieved good performance compared with other baseline sentence vector embedding. From Figure 2.15, Doc2Vec achieved lower error rate of 3.82% compared to other baseline

method. Hence, Doc2Vec is one of the sentence embedding technique will be used in this research.

Model	Error rate
Vector Averaging	10.25%
Bag-of-words	8.10 %
Bag-of-bigrams	7.28 %
Weighted Bag-of-bigrams	5.67%
Paragraph Vector	**3.82%**

Figure 2.15 Performance of Paragraph Vector (Doc2Vec) and other baseline on information retrieval task (Le and Mikolov, 2014)

2. **InferSent**

InferSent is a one of the state-of-art sentence embeddings method that provides semantic sentence representation. Facebook researcher had proposed InterSent in this paper Conneau *et al.* (2017). Figure 2.16 shows the simplified illustration of how InferSent works. InferSent first embed the sentence using sentence encoder. There are several technique in the sentence encoder and one of the most effective technique is Bi-directional LSTM network with max or mean pooling. Each vector is concatenate between forward LSTM and a backward LSTM that able to read sentence in opposite direction. Then, max or mean pooling is used in these concatenated vector to form fixed-length vector.

Natural Language Inference (NLI) or textual entailments is a method of finding directional relationship between text fragment. InferSent make use of Stanford Natural Language Inference (SNLI) dataset and manually labelled with 3 categories (entailment, contradiction and neural), then create NLI classifier. NLI classifier is then used to extract the relations between text and hypothesis.

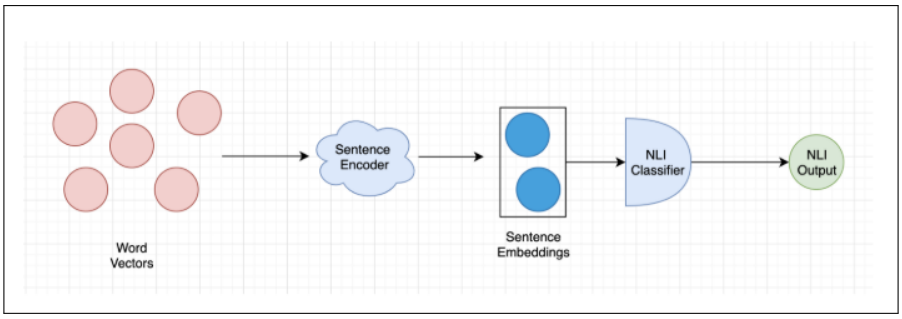


Figure 2.16 Simplified Illustration of InterSent

Due to InferSent's semantic representation, this research will be using InferSent as one of the sentence representation method to be compared with both Doc2Vec and BERT.

3. **BERT**

BERT stands for Bidirection Encoder Representation from Transformers is a state-of-art sentence embedding technique proposed by Google AI language. (Devlin *et al.*, 2018). BERT make use of Transformer, an attention mechanism that learns contextual relation between words in a text.

Figure 2.17 shows the BERT input representation. There are 2 important element in BERT, Masked LM (MLM) and Next Sentence Prediction (NSP). MLM allows bidirectional training where the model uses the context words surrounding a [MASK] token and try to predict what the [MASK] word should be. Besides, in NSP, [CLS] token is inserted at the beginning of first sentence while [SEP] token is insert at the end of each sentence. Combining with [MASK] token, it forms a sequence and the entire sequence will be embeded through Transformer model.

Due to the bi-directional representation behaviour and transformer encoder in BERT, BERT have a highest accuracy in Semantic Text Similarity (STS) score. It obtain the highest (86.5%) similarity score compared to other sentence embedding algorithm such as Pre-OpenAI SOTA (81.0%), BiLSTM+ELmo+Attn (73.3%) and OpenAI GPT (80.0%) (Devlin *et al.*, 2018). Hence, in this research, BERT will be used as one of the sentence embedding method for measuring text semantic similarity

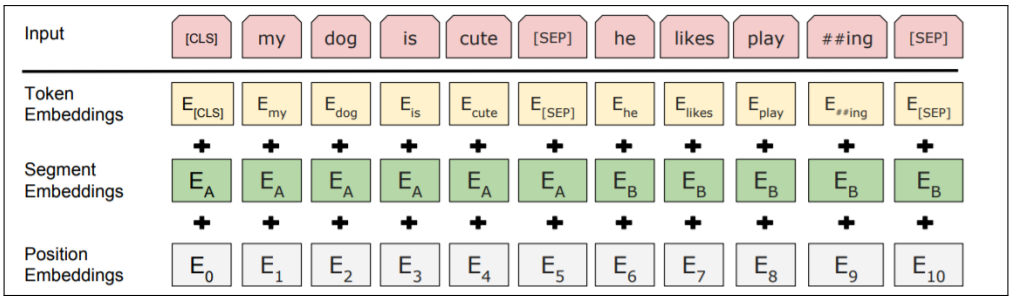


Figure 2.17 BERT input representation (Devlin *et al.*, 2018)

2.5 Prediction Model Technique

Prediction model technique is an algorithm that modelling prediction and provides a guideline throughout the process of building up a predictive model. After obtaining extract keyword from textual data and clustering them based on clustering algorithm and relation extraction, the data is now said to be clean and sanitised. These clean data is then involve in prediction algorithm to build up a prediction model.

2.5.1 Sequential Prediction Algorithm

Sequential prediction is a fast and simple pattern matching algorithm whereby comparing the past data or experience on a linear running timeline. There are many examples for sequential prediction algorithm such as medical condition that occur over time and patient conditions is getting either better or worse based on previous event. Besides, suggestion system that used in music application such as Spotify also involve in sequential prediction.

2.5.2 Neural Network Prediction Algorithm

The biggest advantages of applying neural network in event prediction is the high tolerance and acceptance ability of noisy data and high accuracy. Despite of its complexity and long training time, many researches are more towards to neural network.(Granroth-Wilding and Clark, 2016; Hu *et al.*, 2017; Asencio-Cortés *et al.*, 2017) Neural network use the concept of brain metaphor for information processing and is suitable to discover previously unknown, valid patterns and relationship in large data sets. (Gaur, 2012)

2.5.3 Vector Similarity

Vector similarity is one of the important measure for word vector or sentence vector. Example of technique measuring vector similarity are Cosine Similarity, Euclidean distance, Jaccard distance and word mover's distance. Among these techniques, Cosine similarity is being widely used in applied in NLP domain. (Torabi Asr *et al.*, 2018; Mikolov *et al.*, 2013b; Bekkali and Lachkar, 2019). Equation 2.1 shows the formula of cosine similarity.

$$\frac{A.B}{||A|| * ||B||} \tag{2.1}$$

Cosine similarity is used for chat bot engine (Yang *et al.*, 2018). When user input a random query, the input will be embed into vector and compute cosine similarity calculation with every vector in the model. After that, vector with highest similarity will be selected and returning the original question. The question is said to have the most similar textual semantic to the input query. This question is then linked user to the coresponding answer.

In this research, cosine similarity is one of the most important element to be calculated so that similarity between sentence can be measured for accurate result.

2.6 Existing work of Event Prediction

There are several researchers research on the topic of event prediction based on causality attributes and combining different method on different domain. Table 2.2 shows the existing work of event prediction, focusing on textual data from different domain such as financial, social media, medical etc.

In the previous works, trend of event prediction starts from sequential approach, transforms to deep learning and even involved in neural network model. Researchers found that deep learning and neural network model able to cope with the high variety of unstructured textual data. Besides, event extraction also slowly transform from data-driven approach into knowledge-drive approach which able to observe from the latest work of Dami *et al.* (2018) that using Markov logic network with knowledge-driven

Table 2.2 Existing work on Event Prediction

Author, year	**Work**	**Method**	**Advantages/ Limitation**
(Letham *et al.*, 2013)	Sequential event prediction	Given a sequence past database to predict next event within a current event sequence	Rely on additional user input, supervised ranking algorithm which not suitable for less predictive power past events
(Mirza, 2014)	Temporal sentiment analysis and causal rules extraction from tweets for event prediction	Event prediction based on temporal sentiment and causality from tweets which are from the opinion of people	Sentiment from tweets is uncertain and may affect on the predictive output
(Ding *et al.*, 2015)	Depp learning on event-driven stock market prediction	Extract news text with dense vectors and train with novel neural tensor network, provides prediction based on convolutional neural network	Outcome is more focus on statistical, less focus on financial news that affect the share price in the market
(Hu *et al.*, 2017)	Event Prediction using Compositional Neural Network Model	Use word2vec to represent events and build a compositional neural network model for future event prediction	Multiple event chain in CNN models and limited to learning associations between pairs of events
(Li *et al.*, 2018)	Narrative event evolutionary graph (NEGG) for script event prediction	Extract narrative event chain from newspaper and construct NEEG based on the chain	Require expert knowledge in deep learning to produce NEGG
(Rumi *et al.*, 2018)	Crime event prediction with dynamic features and matrix factorisation	Dynamic features such as visitor entropy, visitor homogeneity, region popularity, visitor ratio and count, observation frequency is taken as measurement in crime event prediction	Different city have different crime data, result may be biased
(Dami *et al.*, 2018)	News events prediction using Markov logic networks	Markov logic network represent complex events by first-order logic and binded with domain-specific causal rules. Web ontology language (OWL) perform causal inference to represent probabilistic knowledge	Domain-specific causal rule must be created precisely before further text processing

extraction approach. In addition, among these previous study, causality is the main attribute that concern by the researchers. In this study, causality will also be the main attribute followed by different embedding technique. Besides, vector similarity will also be used to determine the semantic similarity between corpus embedding and input queries' embedding. Finally, the most similar cause in corpus will be taken and the linked effect will be taken as the possible predicted output.

2.7 Open Issues and Challenges

Event prediction is important and useful for public to have better insight and quick response against event might happen. However, there is several issues and challenges need to be addressed in order to produce a perfect predictive model approaching human-thinking. The issues and challenges are:

1. **Low performance of predictive model due to low resource scenarios**

   For predictive model, it require as much data as possible to train and test its accuracy and improve any weakness that learning from bad data sources. However, one of the problem is that there are scenarios of low resource that the prediction model can obtain. For example, if the model receive data source that are unique and only few events occurs because of these, the predicted output might not be valid and match with human-predicted output.

2. **Word Sense Disambiguation**

   In textual data extraction, word sense disambiguation is an open problem that troubling in determine the sense of word used in a sentence. One of the significant example can be observe in differences between dictionaries. For example, the word "**bank**" has 3 different meaning according to Cambridge Dictionary. First, bank is indicated as financial organisation where people can invest or store money inside, second is sloping raised land along sides of river, third is related to row of similar things. There are 2 sentences, "The **bank** will not be open in Saturdays" and "the river overflowed the **bank**.". The ambiguity between 2 same words but different meaning make machine difficult to detect and extract correct information from its true nature.

3. **Pronoun Resolution**

Pronoun resolution also known as anaphora resolution, is an common problem of resolving what a pronoun, or a noun phrase refers to. (Poesio *et al.*, 2016). Taking example of "John helped Mary", followed by "He was kind". As normal human, we can understand that "he" is indicating to John but not the machine. Machine will automatically define "he" as a new character instead of combining John with "he". Anaphora resolution is still an active research (Choi *et al.*, 2016; Bandaragoda *et al.*, 2018) which contributes to better machine learning technique.

2.8 Summary

In this literature review, concept of event prediction and South China Sea conflict is clearly explained before moving in brief stage of event prediction. Then, stages within event prediction such as text mining, clustering, event relation extraction, prediction algorithm and etc is mentioned with details in the previous subchapter. Finally, this chapter discuss about previous and existing work that done by other researchers on the textual data event prediction in different domain and also the open issues and challenges in event prediction so that reader get better understanding about constraints of event prediction nowadays.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Overview

This study focuses on the implementation of vector similarity measure with different sentence embedding technique for event prediction on South China Sea conflict. In this research, causality relation will be focused. This chapter emphasises the research framework of the study which explains the method and technique used to achieve the objective of the study. This chapter begins with the overview of research framework which consists of three phases and each phase will be briefly explained. Besides, overall research plan will be provided in order to give a better understanding on overall expected result of this study. After that, the chapter continues with brief description of data set which will be used in this study. Finally, the chapter ends with summary.

3.2 Research Framework

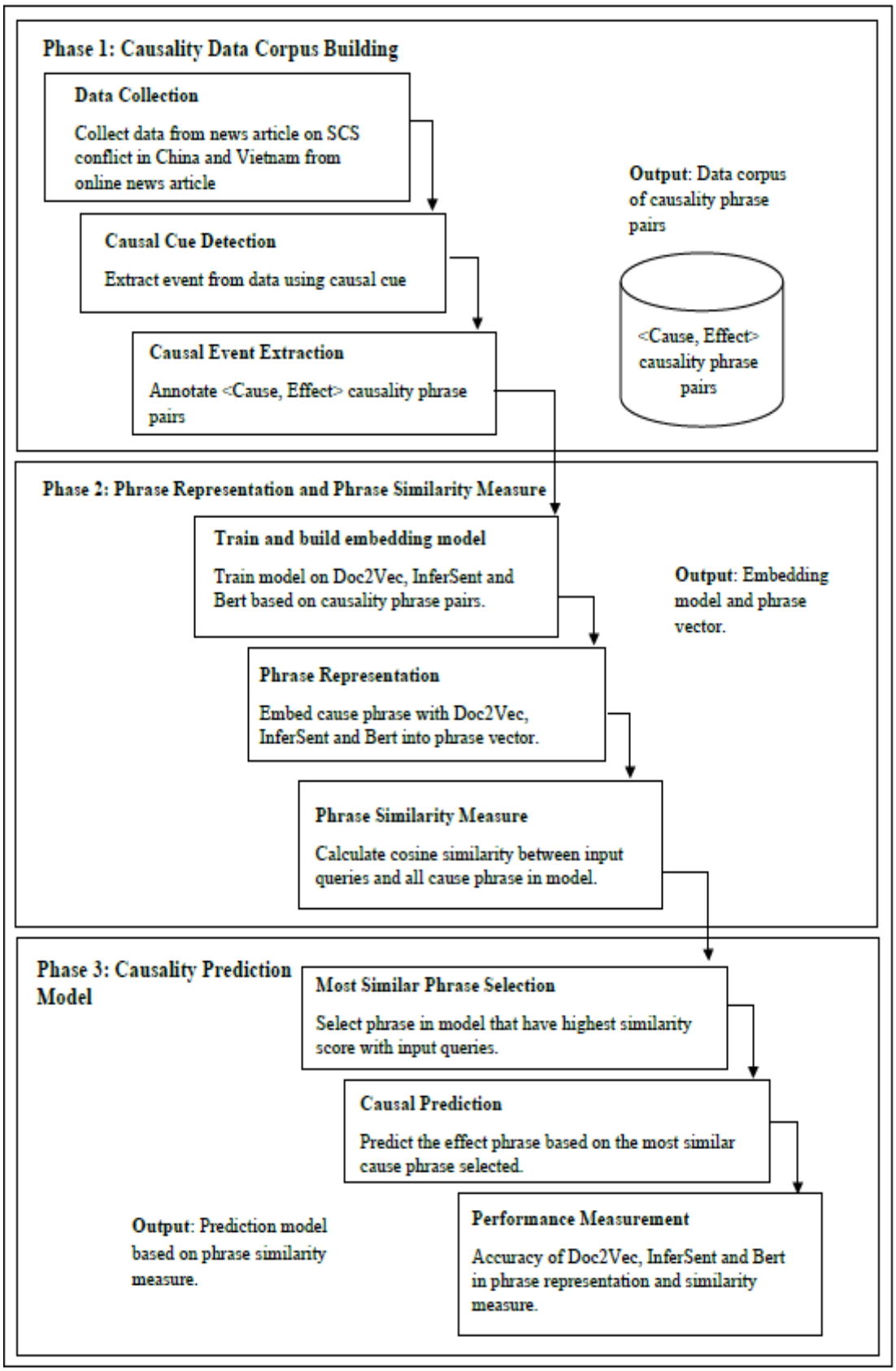


Figure 3.1 Research Framework

The research framework consists of three phases which correspond to the three research objectives in Chapter 1. The three phases are; (i) Causality data corpus building based on South China Sea conflict news article, (ii) Construct abstract causality network using causal event extraction and generalisation, (iii) Causality network embedding model. The flow of these three phases is illustrated on the research framework in Figure 3.1.

In phase 1, news event will be represented in <cause, effect> format where it is named as causality pairs. Data is collected from various online news article that concerned on South China Sea conflict within China and Vietnam. After data collection, the following steps is causality mention extraction. Causality mention extraction will use specific causality connector such as "because", "because of", "after", etc to extract sentences that contain causality event. After that, "cause" and "effect" sentence will be extracted based on the template of <Pattern, Constraints, Priority> in order to clearly identify "cause" and "effect" sentence. Finally, event is represented in the causality pairs format of <Cause, Effect>.

In phase 2, abstract causality network will be generated. <Cause, Effect> causality pairs from phase 1 are further extracted by using verb-and-noun representation. Verb-and-Noun representation is used because it preserve most of the important information rather then losing them. After extracting important information, the words is further generalised using WordNet and VerbNet. Each noun is generalised to its hypernym in WordNet while each verb is generalised to its class in VerbNet. Furthermore, frequently co-occurring word pairs (FCOPA) is used to identify the frequent term and will be represented as node of abstract causality network. The edges of causality network is constructed if there is existing edge in specific event.

In phase 3, prediction model based on causality network embedding model will be proposed. Dual cause-effect transition (Dual-CET) model will be used for adopting essential characteristics of event causality such as asymmetry, transitivity and many-to-many. Besides, a ranking criterion will be used in order to learn the correct event embedding based on true causality pairs. Next, the Dual-CET model

is measured by several evaluation criterion such as Random, Common neighbours, Jaccard's coefficient, etc to measure the effectiveness and reliability of the model.

In addition, the main purpose of overall research plan is to provide a clear idea solve each research question that based on research objective, associated with its methodology and expected result.The overall research plan is tabulated in Table 3.1. In the first objective, the expected result should extracted keywords from South China Sea conflict-related news article in China and Vietnam as well as extract the causality connectors and represent them in causality pairs of <cause, effect> format. Meanwhile, in phase 2, the causality pairs will be further processed with causal event extraction and generalisation and visualise the causality network with nodes and edge based on processed causality pairs. Lastly, in phase 3, the study will focus on proposal of predictive model, which consist of building abstract causality network with frequently co-occurring word pairs (FCOPA) and embed the network with Dual-CET model.

3.2.1 Phase 1

Phase 1 is causality data corpus building and it is important to convert unstructured data from news article to structure form and understandable by the computer. There are several steps to build up a good data corpus which involve specific data collection, causality mention extraction and event representation.

(a) Data Collection

The data is collected from news article based on keyword of South China Sea conflict such as "South China Sea", "ASEAN", "Spratly", "Island", "People's Court",etc. In the news article that consist of these keywords, these articles are taken according to the country where the news being published. This will be further explains in the subsection of data set.

(b) Causality Mention Extraction

In causality mention extraction, the selected news article is then further selected with causality connectors such as "because", "because of", "after", "therefore", "lead to", etc. First of all, sentence segmentation will be used to split all of the sentence in the article. Then, all of the sentence is filtered by causality

Table 3.1 Overall Research Plan

Research Objective	Research Question	Methodology	Result	Performance Measure
To extract <cause, effect> causality pairs from news articles in South China Sea conflict by using causality connectors	How to obtain useful information from news article that related to South China Sea conflict?	• Build news article corpus of SCS conflict from online news article that focus on issues within China and Vietnam • Extract causality pairs of <cause, effect> using causality connectors based on template of <Pattern, Constraints, Priority>	• News article on South China Sea conflict within Vietnam and China is extracted and created news corpus. • Event is extracted and represented in <cause, effect> causality pairs format.	Make sure event pairs is represented in <cause, effect> format
To extract all important information from causality pairs by using verb-noun representation as well as discover general patterns using WordNet and VerbNet.	How to generalise and cluster event extraction and how to relate event pairs with causality?	• Extract all the important information within causality pairs by using verb-noun representation. • Generalise all the verb and noun with WordNet and VerbNet.	• All the important information will be kept while eliminating unused information. • All information is generalised to produce a general, frequent and simple causality patterns.	Visualise causality network with nodes and edge
Train and build an abstract causality network with frequently co-occurring word pairs (FCOPA) and embed the model into a continuous vector space.	How to use the extracted information to train and build a prediction model?	• Build abstract causality network with frequently co-occurring word paris (FCOPA) • Embed causality network with Dual-CET model.	• Prediction model will be generated and able to generate general predicted output based on inputted event. • Accuracy of prediction model is measured.	Examine prediction model by performing several evaluation criterion such as Random, Common neighbours, Jaccard's coefficient.

connectors. After that, regular expression (Regex) will be used to setup extraction rules and pattern and return the result of "cause" and "effect".

(c) Event Representation

In event representation, causality pairs format will be used as each sentence is represented as <*Cause, Effect*>. This representation can be done from obtaining the value based on the pattern of regular expression. For multiple data that obtained in multiple article, all the causality pairs is stored in JSON format.

3.2.2 Phase 2

Phase 2 is discussing about abstract causality network by using causal event extraction and generalisation. After obtaining causality pairs, the pairs still needed to be further processing in order to obtain useful information. At this stage, both cause and effect triple is now in sentence form. In order to obtain and preserve all the information within the triple, verb-and-noun representation is used. After that, all the verb and noun within the triple are generalised using WordNet and VerbNet. Then, the specific causality network is build by using these information. In order to obtain a frequent, general causality pattern, the causality pair is then generalised with frequently co-occurring word pairs (FCOPA) to build up abstract causality network.

(a) Causal Event Extraction

For shrinking down the causality pairs that obtained from the previous steps, causal event extraction is needed. First, POS tagging is used to tag out all the words with their unique characteristics. Then, for each verb, noun and proper noun will be taken and update the triple. Meanwhile, the remaining words will be removed.

(b) Causal Event Generalisation

After extraction, there are still a lot of redundant and similar words within the triple. Hence, generalisation is needed. By using WordNet, noun and proper noun will be generalise to their synonym or hypernym while verb will be generalised to its class by using VerbNet.

(c) Abstract Causality Network

After generalise all the event pairs, the event pairs is linked by using pyvis visualisation library. By using "cause" and "effect" as nodes and each event pair has the edge, build up an abstract causality network. Generalised nouns and verb may be redundant as it is perfect for the abstract causality network in term of diversity.

3.2.3 Phase 3

Phase 3 is to train and build a new embedding model, dual cause-effect transition (Dual-CET) model as the predictive model. In machine learning, a model that generated needs to be trained as the model will dynamically learn the pattern from the inputs and fix any error that occurred during its training. In this study, a new cause event will act as input and predicted effect events will act as output. While adding the input cause event into the model, steps need to be taken as guideline for the overall event prediction.

(a) Dual-CET model

In this study, Dual-CET model is proposed by embedding abstract causality network that created in phase 2. By considering cause-to-effect and effect-to-cause possibilities, a new energy function, f(c,e) is defined,

$$f(c, e) = \|c + t + e\|_1 + \|e + \tau - c\|_1 \tag{3.1}$$

where c and e are cause and effect, +t and $+\tau$ are the encoding of many-to-many, asymmetry and transitivity of event causality.

3.3 Data sets

This study will focus on South China Sea conflict news article from online resource within China and Vietnam. Keywords such as "China", "Vietnam","South China Sea", "Spratly", "World Court" will be used to search online news article. Once the news article matched with the keywords, it is taken as our data sets for further data processing.The collected data is then grouped according to corresponding country based on the location of the news published. For example, if the news are published in Hanoi, the news is grouped into Vietnam since Hanoi is one of the state in Vietnam.

In this study, only 30 online news from China news agency and 30 online news from Vietnam news agency will be taken. In Figure 3.2 and 3.3, sample of news article from Vietnam and China is taken from online resources and put into .txt document file.

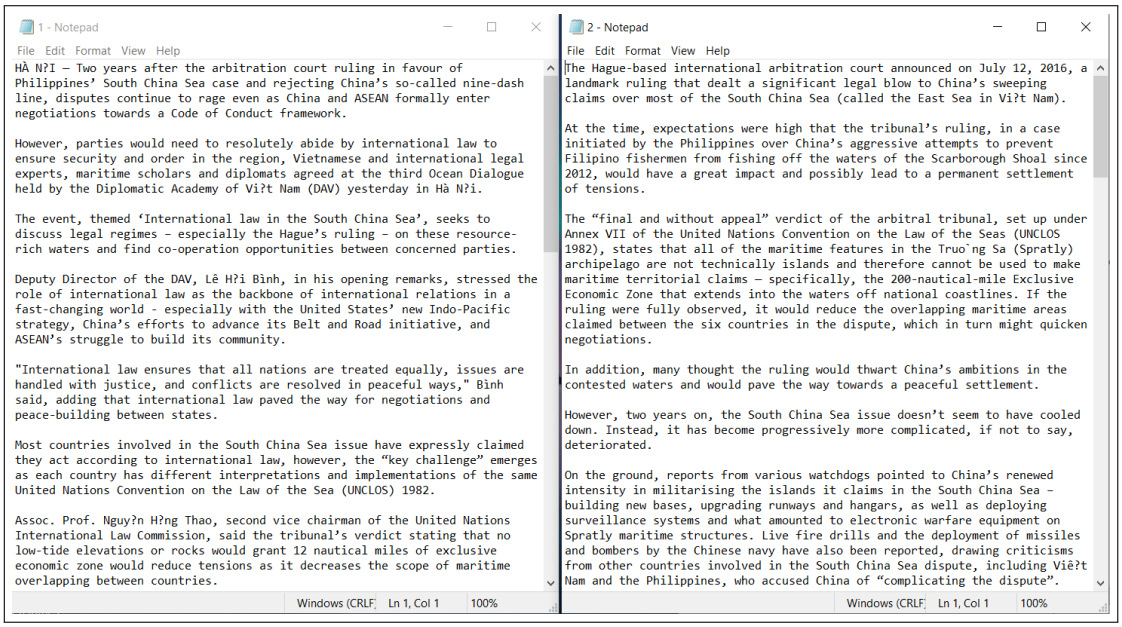


Figure 3.2 Sample of news article from Vietnam New Agency

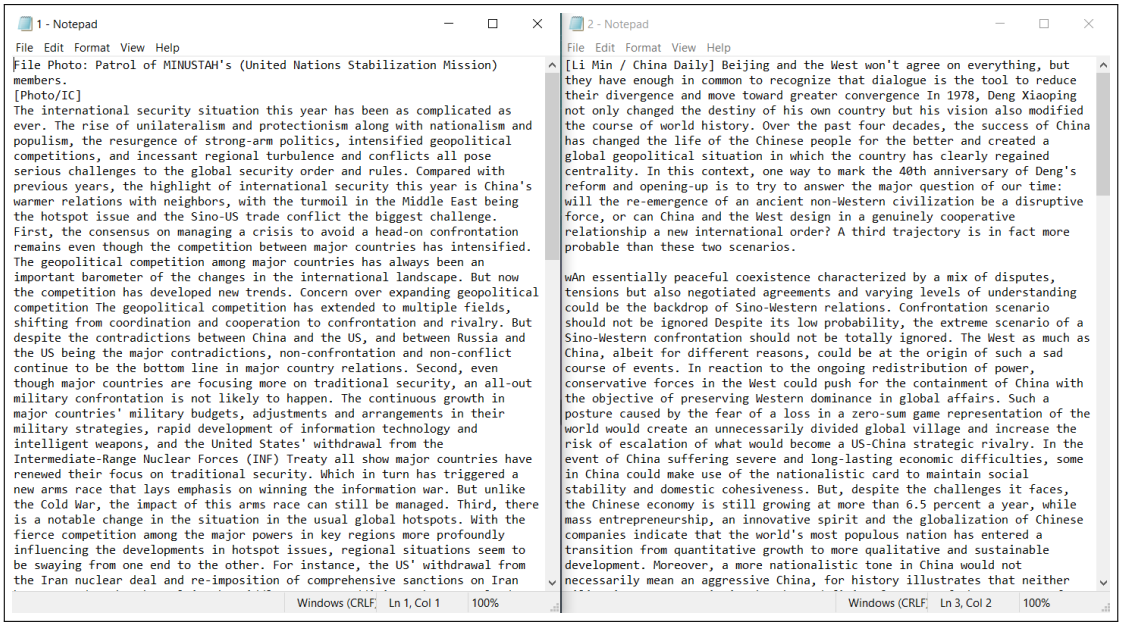


Figure 3.3 Sample of news article form China

To clean the data set, this study will used WebAnno annotation tool to manually annotate keywords based on corresponding entities. After that, stemming process will

be taken to reduce verb into its basic form. Finally, these corresponding entity based on different word characteristics are grouped them into 5 tuple of <Actor, Action, Object, Location, Object>.

3.4 Performance Measurement

To measure the performance of the prediction model, precision and recall as well as F1-score are used. Confusion matrix in Figure 3.4 also provides classification of evaluation between actual event and predicted event.

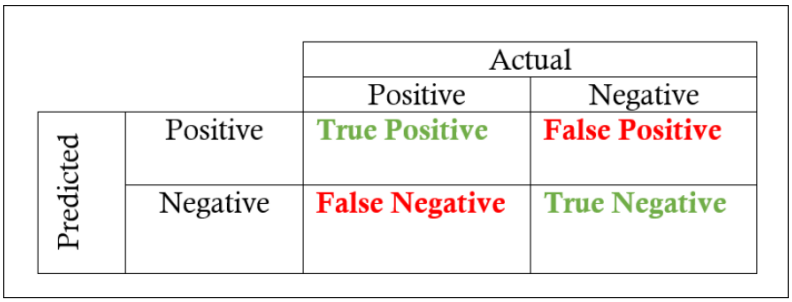


Figure 3.4 Confusion Matrix

Precision is the percentage of relevant results while recall is the percentage of total relevant result that correctly identify from prediction model. Both formula of precision and recall are

$$\textbf{Precision} = \frac{TruePositive}{ActualResults} \tag{3.2}$$

$$\textbf{Recall} = \frac{TruePositive}{PredictedResults} \tag{3.3}$$

To summarise the usage of precision and recall, F1 score is used to give a higher priority to maximizing precision and recall to the prediction model. F1 score is formulated from

$$\textbf{F1 score} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.4}$$

3.5 PSM 1 Gantt Chart

The PSM 1 Gantt Chart is in Appendix A.

3.6 Summary

Chapter 3 discussed the method and steps to use throughout the project. In Research Framework, each of the phase will generate important output and consecutively become input for the next phases. For phase 1, data corpus is built from unstructure textual data from SCS-related news article and represent in 5 tuple format. Then, in phase 2, the sanitised event are generalised and an abstraction tree is built by using HAC hierarchical clustering. Causality prediction rules is also generated in order to generate predicted effect events. Finally, in phase 3, prediction model is trained and the output is filtered in order to produce high quality and reliability output. The prediction model is then evaluate by performance measurement such s precision and recall. Data set used in this study are also stated as it is the main source for this project.

CHAPTER 4

EXPERIMENTAL SETUP

4.1 Introduction

In this chapter, an experimental setup is done by working on phase 1 in the Research Framework. Steps in Phase 1 such as data collection, text preprocessing and event representation will be explained in details. After that, this chapter displays the initial result and finding of phase 1. Finally, the chapter ends with a summary of the chapter.

4.2 Text Preprocessing and Event Extraction

In phase 1, the data is collected from online news article that is related to South China Sea conflict with the selected keywords. In order to present the process of text preprocessing, sample sentence is taken from one of the news article. The sample of sentences is *"Two years after the arbitration court ruling in favour of Philippines' South China Sea case and rejecting China's so-called nine-dash line, disputes continue to rage even as China and ASEAN formally enter negotiations towards a Code of Conduct framework."*, taken from VietnamNews. In this sentence, we can obtain that the SCS disputes continue to rise as issues occur in SCS countries.

In phase 1, text preprocessing will be the primary work. Purpose of text preprocessing is to structure the unstructured textual data and extract useful information while ignoring those duplicated and meaningless words. There are few steps involved in text preprocessing such as annotation, stemming and representation. Figure 4.1 shows the flow of text preprocessing in phase 1.

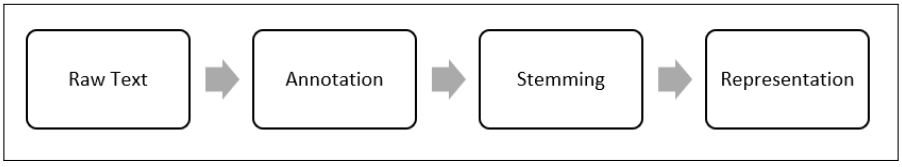


Figure 4.1 Flow of Text Preprocessing

At the beginning of the project, this project tend to extract information by using standard Named Entity Recognisation (NER). NER is a part of information extraction that identify named entity within a sentence. However, standard NER has its limitation that only able to recognise person, organisation, location and time. Figure 4.2 shows the limitation of NER extraction by using sample sentence.

Figure 4.2 Limitation of standard NER for only extracting person, organisation, location and time

Thus, manual annotation is used to annotate important entity such as action, object and instrument. For annotation, WebAnno annotation tool is used. WebAnno is a general purpose web-based linguistic annotation tool including various layers of morphological, syntactical and semantic annotations. By using WebAnno, this study is able to add custom entity and annotate them with matched words. Figure 4.3 shows annotation of sample sentence in WebAnno.

Figure 4.3 Annotation of sample sentence in WebAnno

After annotation, the output is printed using CoNLL-2002 format. CoNLL-2002 format is a data output format which consists of two columns separated by a single space. There are 2 attributes in CoNLL-2002. The first one is tag format which B denotes first item and I denotes non-initial words. The second one is the entity tag of

the words. Figure 4.4 shows the CoNLL-2002 data output of annotation. In this format, the custom entity such as *TIME*, *ACTION* and *OBJECT* will be defined accordingly.

```
Two B-TIME
years I-TIME
after O
the O
arbitration B-ORG
court I-ORG
ruling B-ACTION
in O
favour O
of O
Philippines B-LOC
' O
South B-LOC
China I-LOC
Sea I-LOC
case B-OBJECT
and O
rejecting B-ACTION
China B-LOC
' O
s O
so-called O
nine-dash B-OBJECT
line I-OBJECT
, O
disputes B-OBJECT
continue O
to O
rage B-ACTION
even O
as O
China B-LOC
and O
ASEAN B-LOC
formally O
enter B-ACTION
negotiations I-ACTION
towards O
a O
Code B-OBJECT
of I-OBJECT
Conduct I-OBJECT
framework I-OBJECT
. I-OBJECT
```

Figure 4.4 Annotation in CoNLL-2002 format

After that, all the entity tag with *ACTION* will be normalised through stemming. Stemming is the process where convert verb into its basic form. For example, in the sample sentence, "ruling" and "rejecting" are the entity tag with *ACTION*. After stemming, they will become "rule" and "reject" respectively. In the following steps, these custom entity will be assigned into 5 tuple of <Actor, Action, Object, Location, Timestamp> for event representation. Both stemming and representation will be done in the future.

4.3 Summary

In this chapter, the preliminary works in phase 1 is done on the data collected from South China conflict.First, the annotation model is preset with custom entity tag because there are limitation of entity in standard NER. The limited entity in standard NER are person, organistion, location and time which are required to have more in this study. Then, the annotation is then output as CoNLL-2002 format as it provides simple yet essential output for the further text processing. After that, the words with custom entity of *ACTION* are stemmed into its basic form. Finally, these customer NER will be represented in 5 tuple approach. These preliminary works is important so that the following text processing will be successful and works on the right path.

CHAPTER 5

CONCLUSION

5.1 Conclusion Remarks

After performing preliminary task in previous chapter, the initial result had been partially achieved. Data source is cleaned and sanitised. Cleaned data is suitable for further processing to obtain valuable information. However, there are a lot of works needs to be done in order to reach the stage of building up prediction model and it will be discussed in the next section. In this study, the main attributes is causality. Causality is used to indicate the relation between cause and effect. In the future works, this study needs to extract the causality from news article and compare with world ontology to get precise knowledge on the same domain. Besides, Pundit algorithm needs to be followed from step to step as a guideline in this study. Pundit algorithm provides a prediction framework that start from learning the causality of events until the prediction of future event. By obeying Pundit algorithm, the predicted effect event will have high reliability and precision.

5.2 Future Works

After annotating entity from sample sentence in previous chapter, these entities need to be assigned based on their characteristic which manually defined by researcher. In addition, another challenge is to put all these entities into corresponding tuple <Actor, Action, Object, Location, Timestamp>. This step might require some recursive algorithm that constantly load matched entity into the tuple.

In the next phase, which is phase 2, the extracted tuples are being generalised in order to reduce the size of corpus. Generalisation is measured by using the similarity of the event pairs based on world ontology. By generalising the event pairs, multiple generalisation path will be created and the shortest path are defined as minimal generalisation path. With minimal generalisation path, an abstraction tree is built and the nodes within the abstraction tree is cluster through HAC hierarchical clustering.

After that, causality prediction rules is set in the form of <Pattern, Constraints, Priority> for the preparation of input cause event to generate most possible effect event.

In phase 3, the prediction model based on abstraction tree is built. Test data are taken from new input cause event and the new event will propagate throughout the abstraction tree to find the matched node. During the propagation, minimal generalisation path and causality prediction rules are used to guide the input events for the desired effect event. To maximize the performance of prediction model, filtering process will take place to eliminate inappropriate predicted effect event. Filtering process will be handled by PMCI calculation to match up the relativeness between cause and effect. Finally, performance measure by using precision and recall will be evaluated the accuracy and precision of the prediction model.

5.3 Summary

This chapter discusses the overall conclusion and the future works that require completing within this project. After finishing the steps in previous chapter, there are still a lot of works need to be done, such as event generalisation, causality prediction rules generation, building up an abstraction tree, etc. Future work will be done on the following chapter and causality will be focus in this study so that event pairs has strong causal relation that satisfy the requirement of event prediction.

REFERENCES

Afzal, M. T., Maurer, H., Balke, W.-T. and Kulathuramaiyer, N. (2010). Rule based autonomous citation mining with TIERL. *Journal of Digital Information Management*. 8(3), 196–204.

Asencio-Cortés, G., Martínez-Álvarez, F., Troncoso, A. and Morales-Esteban, A. (2017). Medium–large earthquake magnitude prediction in Tokyo with artificial neural networks. *Neural Computing and Applications*. 28(5), 1043–1055.

Asim, K., Martínez-Álvarez, F., Basit, A. and Iqbal, T. (2017). Earthquake magnitude prediction in Hindukush region using machine learning techniques. *Natural Hazards*. 85(1), 471–486.

Baker, J. (2019). *Machine Learning Versus The News*. Retrievable at `https://towardsdatascience.com/machine-learning-versus-the-news-3b5b479d8e6a`.

Bandaragoda, T. R., De Silva, D., Alahakoon, D., Ranasinghe, W. and Bolton, D. (2018). Text Mining for Personalized Knowledge Extraction From Online Support Groups. *Journal of the Association for Information Science and Technology*. 69(12), 1446–1459.

Battistella, G. (2005). The World's Biggest Propaganda Agency. *Reporters Without Borders*.

Bekkali, M. and Lachkar, A. (2019). An effective short text conceptualization based on new short text similarity. *Social Network Analysis and Mining*. 9(1), 1.

Bell, B. (2016). *The Road to Prediction: Using Unstructured Data*. Retrievable at `https://www.dataversity.net/the-road-to-prediction-using-unstructured-data/`.

Bhardwaj, B. (2016). Text mining, its utilities, challenges and clustering techniques. *International Journal of Computer Applications*. 135(7), 975–8887.

Casert, R. (2016). *Belgium ramps up security for lone suspect in Jewish Museum attack*. Retrievable at `https://archive.is/20140527190719/http://www.theglobeandmail.com/news/world/`

police-hunt-brussels-jewish-museum-shooter-france-tightens-security/article18835439/#selection-583.1-583.67.

Chen, X., Wang, S., Tang, Y. and Hao, T. (2019). A bibliometric analysis of event detection in social media. *Online Information Review*.

ChinaPower (2016). *How much trade transits the South China Sea?* Retrievable at https://chinapower.csis.org/much-trade-transits-south-china-sea/.

Choi, M., Liu, H., Baumgartner, W., Zobel, J. and Verspoor, K. (2016). Coreference resolution improves extraction of Biological Expression Language statements from texts. *Database*. 2016.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L. and Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Dami, S., Barforoush, A. A. and Shirazi, H. (2018). News events prediction using Markov logic networks. *Journal of Information Science*. 44(1), 91–109.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J. and Manning, C. D. (????). Universal Dependencies: A cross-linguistic typology.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

D.Goltz, J. (2018). *Earthquake advisories can save lives*. Retrievable at https://www.japantimes.co.jp/opinion/2018/04/17/commentary/japan-commentary/earthquake-advisories-can-save-lives/#.XIc8d7NbzeQ.

Ding, X., Zhang, Y., Liu, T. and Duan, J. (2015). Deep learning for event-driven stock prediction. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

E.Hayes, D. (1980). South China Sea, PACIFIC OCEAN. *The Tectonic and Geologic Evolution of Southeast Asian Seas and Islands*.

Erevelles, S., Fukawa, N. and Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. *Journal of Business Research*. 69(2), 897–904.

Fedoryszak, M., Frederick, B., Rajaram, V. and Zhong, C. (2019). Real-time event detection on social data streams. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2774–2782.

Fensom, A. (2016). $5 Trillion Meltdown - What If China Shuts Down the South China Sea? *The National Interest*. Retrievable at https://nationalinterest.org/blog/5-trillion-meltdown-what-if-china-shuts-down-the-south-china-16996.

Frunza, M.-C. (2016). Chapter 3B - Exploring Unstructured Data. In Frunza, M.-C. (Ed.) *Solving Modern Crime in Financial Markets*. (pp. 263 – 273). Academic Press. ISBN 978-0-12-804494-0. doi:https://doi.org/10.1016/B978-0-12-804494-0.00019-X. Retrievable at http://www.sciencedirect.com/science/article/pii/B978012804494000019X.

Gaur, P. (2012). Neural networks in data mining. *International Journal of Electronics and Computer Science Engineering (IJECSE, ISSN: 2277-1956)*. 1(03), 1449–1453.

Gonzales, R. (2014). The Spratly Islands Dispute: International Law, Conflicting Claims, and Alternative Frameworks For Dispute Resolution.

Granroth-Wilding, M. and Clark, S. (2016). What happens next? event prediction using a compositional neural network model. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Grolinger, K., L'Heureux, A., Capretz, M. A. and Seewald, L. (2016). Energy forecasting for event venues: Big data and prediction accuracy. *Energy and Buildings*. 112, 222–233.

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L. and Szpakowicz, S. (2010). SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. July. Uppsala, Sweden: Association for Computational Linguistics, 33–38. Retrievable at https://www.aclweb.org/anthology/S10-1006.

Hogenboom, F., Frasincar, F., Kaymak, U. and De Jong, F. (2011). An overview of event extraction from text. In *Workshop on Detection, Representation, and*

Exploitation of Events in the Semantic Web (DeRiVE 2011) at Tenth International Semantic Web Conference (ISWC 2011), vol. 779. Citeseer, 48–57.

Hu, L., Li, J., Nie, L., Li, X.-L. and Shao, C. (2017). What happens next? Future subevent prediction using contextual hierarchical LSTM. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Kulkarni, R. (2019). *Big Data Goes Big*. Retrievable at `https://www.forbes.com/sites/rkulkarni/2019/02/07/big-data-goes-big/#5c7be26420d7`.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*. 1188–1196.

Letham, B., Rudin, C. and Madigan, D. (2013). Sequential event prediction. *Machine learning*. 93(2-3), 357–380.

Li, Z., Ding, X. and Liu, T. (2018). Constructing narrative event evolutionary graph for script event prediction. *arXiv preprint arXiv:1805.05081*.

Lilleberg, J., Zhu, Y. and Zhang, Y. (2015). Support vector machines and word2vec for text classification with semantic features. In *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*. IEEE, 136–140.

Lima, E., Mues, C. and Baesens, B. (2009). Domain knowledge integration in data mining using decision tables: Case studies in churn prediction. *Journal of the Operational Research Society*. 60(8), 1096–1106.

Maziarz, M., Heagerty, P., Cai, T. and Zheng, Y. (2017). On longitudinal prediction with time-to-event outcome: Comparison of modeling options. *Biometrics*. 73(1), 83–93.

Mele, I., Bahrainian, S. A. and Crestani, F. (2019). Event mining and timeliness analysis from heterogeneous news streams. *Information Processing & Management*. 56(3), 969–993.

Michael Benz, M. M. *(2017)*.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

Mirza, P. (2014). Extracting temporal and causal relations between events. In *Proceedings of the ACL 2014 Student Research Workshop*. 10–17.

mongoDB (2016). *Unstructured Data In Big Data*. Retrievable at `https://www.mongodb.com/scale/unstructured-data-in-big-data`.

Mulia, I. E., Asano, T. and Nagayama, A. (2016). Real-time forecasting of near-field tsunami waveforms at coastal areas using a regularized extreme learning machine. *Coastal Engineering*. 109, 1–8.

news, B. (2014). *Vietnam boat sinks after collision with Chinese vessel*. Retrievable at `https://www.bbc.com/news/world-asia-27583564`.

Ning, Y., Muthiah, S., Rangwala, H. and Ramakrishnan, N. (2016). Modeling precursors for event forecasting via nested multi-instance learning. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1095–1104.

Pechsiri, C. and Piriyakul, R. (2010). Explanation knowledge graph construction through causality extraction from texts. *Journal of computer science and technology*. 25(5), 1055–1070.

Pennington, J., Socher, R. and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

Poesio, M., Stuckardt, R. and Versley, Y. (2016). *Anaphora resolution: Algorithms, resources, and applications*. Springer.

Preethi, P. G., Uma, V. *et al.* (2015). Temporal sentiment analysis and causal rules extraction from tweets for event prediction. *Procedia computer science*. 48, 84–89.

Radinsky, K., Davidovich, S. and Markovitch, S. (2012). Learning causality for news events prediction. In *Proceedings of the 21st international conference on World Wide Web*. ACM, 909–918.

Riloff, E. *et al.* (1993). Automatically constructing a dictionary for information extraction tasks. In *AAAI*, vol. 1. Citeseer, 2–1.

Ritter, A., Etzioni, O. and Clark, S. (2012). Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1104–1112.

Rospocher, M., van Erp, M., Vossen, P., Fokkens, A., Aldabe, I., Rigau, G., Soroa, A., Ploeger, T. and Bogaard, T. (2016). Building event-centric knowledge graphs from news. *Journal of Web Semantics*. 37, 132–151.

Rumi, S. K., Deng, K. and Salim, F. D. (2018). Crime event prediction with dynamic features. *EPJ Data Science*. 7(1), 43.

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D. and Sperling, J. (2009). Twitterstand: news in tweets. In *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems*. 42–51.

Schinas, M., Papadopoulos, S., Kompatsiaris, Y. and Mitkas, P. (2018). Event detection and retrieval on social media. *arXiv preprint arXiv:1807.03675*.

Tan, A.-H. *et al.* (1999). Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, vol. 8. sn, 65–70.

Taylor, C. (2018). *Structured vs. Unstructured Data*. Retrievable at `https://www.datamation.com/big-data/structured-vs-unstructured-data.html`.

Torabi Asr, F., Zinkov, R. and Jones, M. (2018). Querying Word Embeddings for Similarity and Relatedness. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. June. New Orleans, Louisiana: Association for Computational Linguistics, 675–684. doi:10.18653/v1/N18-1062. Retrievable at `https://www.aclweb.org/anthology/N18-1062`.

Tung, K.-C., Wang, E. T. and Chen, A. L. (2016). Mining event sequences from social media for election prediction. In *Industrial Conference on Data Mining*. Springer, 266–281.

Valenzuela-Escárcega, M. A., Hahn-Powell, G., Surdeanu, M. and Hicks, T. (2015). A domain-independent rule-based framework for event extraction. *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, 127–132.

Vargas, M. R., De Lima, B. S. and Evsukoff, A. G. (2017). Deep learning for stock market prediction from financial news articles. In *2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*. IEEE, 60–65.

Vossen, P., Agerri, R., Aldabe, I., Cybulska, A., van Erp, M., Fokkens, A., Laparra, E., Minard, A.-L., Aprosio, A. P., Rigau, G. *et al.* (2016). NewsReader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems*. 110, 60–85.

Wu, Y., Sun, H. and Yan, C. (2017). An event timeline extraction method based on news corpus. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)(*. IEEE, 697–702.

Xiang, W. and Wang, B. (2019). A Survey of Event Extraction From Text. *IEEE Access*. 7, 173111–173137.

Yang, Y., Yuan, S., Cer, D., Kong, S.-Y., Constant, N., Pilar, P., Ge, H., Sung, Y.-H., Strope, B. and Kurzweil, R. (2018). Learning semantic textual similarity from conversations. *arXiv preprint arXiv:1804.07754*.

Yzaguirre, A., Smit, M. and Warren, R. (2016). Newspaper archives+ text mining= rich sources of historical geo-spatial data. In *IOP Conference Series: Earth and Environmental Science*, vol. 34. IOP Publishing, 012043.

Zhang, D., Xu, H., Su, Z. and Xu, Y. (2015). Chinese comments sentiment classification based on word2vec and SVMperf. *Expert Systems with Applications*. 42(4), 1857–1863.

Zhao, S., Wang, Q., Massung, S., Qin, B., Liu, T., Wang, B. and Zhai, C. (2017). Constructing and embedding abstract event causality networks from text snippets. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 335–344.

Zhen, L. (2014). What's China's 'Nine-Dash Line' and why has it created so much tension in the South China Sea? *South China Morning Post*.

Appendix A PSM Gantt Chart