# CHAPTER 1

# INTRODUCTION

> "..., research in psychology (for example, Petre(1995)) shows that in many cases a considerable amount of expertise and background knowledge is needed to interpret a graphic correctly, and novices may be better off with a textual representation of the information."
>
> – *Ehud Reiter & Robert Dale, in [1]*

Recently, in Indonesia, there are diseases which are not handled properly, especially communicable diseases. This case occurs because of the limited awareness and preparedness in confronting disease threats. Therefore, since an early warning system is indispensable by Indonesian community, then the method for informing the information related to disease threats is proposed in this study.

This chapter discusses the rationale in Section 1.1 that explains the background of this study and the related problem situation. Theories and concepts used to conceptualize this study are discussed in Section 1.2, while Section 1.3 discusses the variables related to the problem and their relationship to the paradigm of this study. The intended problem within this study is explained in Section 1.4. Section 1.5 discusses the proposed approach for solving the intended problem. Besides, this study describes some assumptions in Section 1.6, while Section 1.7 describes the scope of works and delimitations. Finally, the contributions of this study are described in Section 1.8.

## 1.1 Rationale

Nowadays, the early warning system in Indonesia is built based on health surveillance data. Health surveillance data is often presented in a graphical representation, such as line charts. Indonesian communities in rural areas still does not have enough capabilities to read and understand the intended charts [2]. Usually, the data is often interpreted manually by experts. In fact, experts require 2-4 days [3] for interpreting health surveillance data. Meanwhile, experts who are able to interpret those data are not always available [4] in rural areas. Therefore, a system for interpreting the chart is needed, so that the community can understand the information represented by the related chart. By understanding the intended information, the awareness and preparedness of the community against the disease threat can be improved.

In the last few years, there are studies developed for the intended system. In 2008, Demir et al. [5] discussed the novel approach for generating a brief textual summary of a simple bar chart into natural language –English. Their proposed system used a logical representation of the graphic's core message produced by SIGHT system and the XML representation, where the XML representation represented chart components. The generated summary of the bar chart content was unique because of there was no domain restriction. In 2014, Moraes et al. [6] presented the Natural Language Generation (NLG) system which was able to generate summaries of simple line chart. This study considered the reading level in generating the intended summaries. Similar to Demir et al. [5], Moraes et al. [6] proposed system generated summaries in English. Besides, in the same year, Mahamood et al. [7] redeveloped the Arria engine, so that this engine was not only able to generate the textual representation, but also the annotated graphs. In other words, these representations are known as multi-modal document.

In Indonesia, the study conducted by Pratomo and Barmawi [2] in 2013 has successfully developed the health surveillance chart interpreter system based on Indonesian language. This system generated health surveillance summaries using simple line chart features, such as the chart title, axis titles, legends, and data values. Based on the proposed evaluation, the generated summaries is quite good natural for readers. However, the naturalness of generated summaries is still limited, so that the probability of mis-perception in interpreting the related chart is increased. This is caused by the naturalness of sentences that influence the readers' understanding [8]. Other weaknesses of the system proposed by Pratomo and Barmawi [2] is that it requires the manual input from the experts to determine the proper interpretation words and it cannot handle the multi-lines chart.

Since the automatic chart interpreter system is useful for Indonesian community, then this study is intended to improve the system developed by Pratomo [2]. This study is conducted based on the limitations of Pratomo's system [2].

## 1.2 Theoretical Framework

In decades, many studies about textual generation have been applied on several domain areas and disciplines, regardless whether the generated texts are in the forms of summaries or report. Most of them apply the Natural Language Generation (NLG) technology. In NLG technology, there are several approaches classified as the inflexible methods and flexible methods. They are canned text, template-based, phrase-based, and feature-based [9][10][11].

Canned text is a simple method classified as the inflexible method because it generates sentences based on predefined sentences without any change [12]. It requires many predefined sentences if generating various sentence combinations are lined, this is considered to be wasteful and costly, for example, error and warning messages. At a glance, template-based is not different from mail-merge. The system generates sentences using language patterns through a fill-in-the black mechanism [12][13]. This technique is often used by domains which do not need many sentence combinations, such as the health surveillance summaries. It is more flexible rather than canned text because the complex linguistic knowledge is not required.

The phrase-based and feature-based are sophisticated techniques because they have considered the sentence level (i.e grammar rule) and the discourse level (i.e text plan). Although they are difficult to build, they are powerful and robustbecause it requires linguistic knowledge [10][13]. For deciding the proper technique, they should consider the requirement of each domain [1][12].

## 1.3 Conceptual Framework

The basic concept of the proposed method is generating summary of health surveillance chart in Indonesian language. To achieve this objective, natural language generation method is proposed. Since the intended summary is represented as contextual representation, then semi template-based method is used in this method. **Instead of using template-based method, semi template-based method uses a number of possible sentence patterns generated based on the corpus of health surveillance summaries**. These patterns are called as the template. The template consists of codes of context and attribute. Context code refers to the context conveyed by a sentence, while attribute code is "blank slots", where "blank slots" is chunks of sentence which can be replaced by a particular variable value based on the context of the conveyed sentence. Furthermore, the system selects one template randomly. The selected template contains some attribute code that should be filled with entities of attributes (e.g. Malaria, rainfall, 2001 - 2002 , or spike) for realizing the final sentence.

Since the naturalness of generated language is considered for generating natural language, then the naturalness should be evaluated. **In order to evaluate the language**

**naturalness, clarity, readability, and general appropriateness are considered** [14]. The clarity is closely related to the reader's understanding. The readability refers to the sentence relationship between one and another. Meanwhile, the general appropriateness refers to whether the sentences can help reader to enrich the knowledge about the public health surveillance. Those three aspects are evaluated by the human-based evaluation that involves novices and experts.

## 1.4 Problem Statements

Based on the rationale of this research in Section 1.1, Indonesian communities requires the automatic chart interpreter system for generating the textual representation. A method for automatic chart interpreter system has been developed by Pratomo and Barmawi [2] in 2013. However, **the naturalness of Pratomo's system is still limited**. Besides, Partomo's system has several other problems, such as **it cannot handle two-lines chart** and **it still depends on the expert's input**. Therefore, this study proposed a method for reaching the naturalness and improving other related problems.

## 1.5 Hypothesis

Based on problems of Pratomo's system [2], the proposed method applies semi template-based method for improving the naturalness of summaries generated by the system. Summaries generated by the system based on semi template can ensure the achievement of the intended naturalness. **This condition can be reached because the generated templates are built based on the corpus of the existing health surveillance summaries, so that the generated sentences are more natural**.

In order to improve the performance of Pratomo's system in the case of two-lines chart and avoiding the expert's input, this study proposed a natural language generation (NLG) system based on Data-to-Text architecture [15]. These problems can be solved by the intended architecture because **this architecture contains Data Interpretation which is useful for selecting data which has relationship with the conveyed event and useful for inferencing the intended event into a particular interpretation word**. Furthermore, the selected data and data inference are used to generate the intended summaries using semi template-based method.

## 1.6 Assumption

This study assumes that the input to the proposed system is two-lines chart table representing to two-lines chart features, such as the chart title, axis name, legends, and data values. These features refer to chart features that have been extracted by Pratomo and Nugroho in [16]. They can be represented by users (i.e novices and experts) in the spreadsheet document using a particular table structure. The description about types of table structure and how to represent the intended features into the related table structure is provided by the proposed system. Meanwhile, the related chart features apply the general standard in health surveillance (i.e early warning system), where data values are based on the weekly data.

Human-written sentence patterns called as template are constructed based on the examples of health surveillance summaries collected from various sources, such as articles, reports, and bulletins in the intended domain. These examples of health surveillance summaries are assumed that their sentence patterns have been constructed by humans (i.e experts) in the well-formed structure of Indonesian language.

## 1.7 Scope and Delimitation

In order to generate the summaries of chart image into an Indonesian language, this study formulated the scope and delimitation as follows :

1. Data used in this study is health surveillance data about the communicable diseases or diseases that potentially give rise to outbreak.
2. These data consist of two dependent variables namely: the number of cases and the influence factor related to weather.
3. These data and other chart features are represented into two-lines chart table.
4. This study does not handle data related to the prevalence of cases and the incident rate of cases.
5. This study does not handle the fluctuation data and the lacking pattern that may be occurred in public health surveillance area.
6. Sentences to be conveyed in the health surveillance summaries are sentences describing the chart overview, data trends, extreme data, sudden change of data, and outbreak situation.

## 1.8 Importance of The Study

This study aims to improve the performance of Pratomo's system by proposing the automatic chart interpreter system based on Data-to-Text architecture for generating the health surveillance summaries using semi template-based method.

This system is aimed at assisting Indonesian communities (i.e novices) in understanding data represented by two-lines chart, so that the community awareness and preparedness can be improved. This is one of support components to improve the national health that in line with the goal of Indonesian government.