CHAPTER 1: INTRODUCTION

1.1 Rationale

Chart interpretation has an important role in many aspects. For example, in a research, reading and interpreting a chart will give significant impact, not only to take a comment or giving a response, but also important for taking the critical decision. Chart interpretation also has an important role in the health care. Many decisions, judgments, or conclusions have taken based on any chart interpretation.

However, nowadays chart is still a problem because the charts are normally interpreted manually. In rural areas, often there are no professional available to interpret charts. The non-expert person, usually those who have non-engineering education background, often interpret charts incorrectly. This misinterpretation can lead to a bigger problem while the conclusion is far from the fact because of many information which is gained to support a decision is incomplete or inaccurate [1].

Considering the important role of chart interpretation and also from the exploration so far that none of automatic chart interpretation based on Indonesian language, then this thesis entitled *Automatic Chart Interpretation with an Application in Health Surveillance* conducted. This thesis will discuss on an automatic chart interpretation system, designed and developed in Indonesian language, and displayed in most natural language so that any people in Indonesia can interpret the result easily. According to the title of this thesis, the chart is limited to the health related chart. The sample of the chart is illustrated in Figure 1.

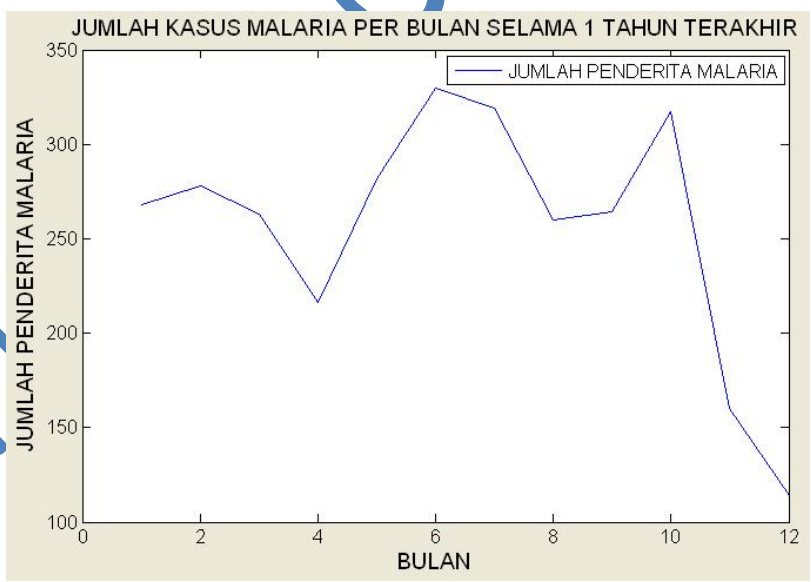

Figure 1 - Chart Example

1.2 The Problem

Chart interpretation nowadays is still a problem, especially to a people with non-engineering education background. Interpretation of a chart is often to be very limited when a given chart is actually full of meaning. This condition is very unfortunate because the reader is unable to understand what the content is actually conveyed by the chart.Incorrect interpretation on the chart will give the wrong perception of the given data. This mistake certainly has significant impact on the decision to be taken based on that chart interpretation.

The problems also come from the Natural Language Processing methods where the existing NLG systems used very large size of the corpus as a language direction. In addition, there are no grammatical rules used in existing NLG systems. The details of this thesis problem are as follows:

1. How can the chart image be read automatically?
2. How can a NLG system with a small corpus be built?

1.3 Theoretical Framework

Given an image such as depicted in Figure 1, then the information to be extracted from the image would be represented by collection of pixels. Certain information can be extracted by examining the layout and structure of the pixels. Symbols are represented by character-like structure while the curve is represented by un-connected layout of pixels. The extraction tools should pay attention on this specific structure and layout and hence appropriate tools should be carefully selected. Much of the algorithms developed for this kind of task fall into the area of Image Processing. To extract character-like symbols, OCR can be used while in extracting curve, certain pixel detection tools should be developed.

If the symbols and values information representing the curve have been collected, a report can be formulated from them. Equipped with appropriate specialty, human being can perform such task without any difficulty. However, the same task would be very difficult to be done by computer, especially when the report has the same quality as those written by a human being. To generate such report, some aspects must be considered such as the grammar, dialect, criticality of the information, etc. Works on these problems fall into the area of Natural Language Processing. To generate a title, certain type of words are selected from the available information, while generating curve analysis result, some processes such as chart value extraction, pattern detection, etc should be done first.

1.4 Conceptual Framework/Paradigm

The main tools used in this thesis consist of Image Processing and Natural Language Processing parts. The image processing part is used to extract the features from the chart image. The features extracted from the chart include chart title, chart legend, axis title, axis number, and the pairs of (*x, y*) number representing by the chart line. The text features such as title, legend, axis number, and axis title, are extracted by Optical Character Recognition (OCR) process. The next process is extracting the line coordinate after segmenting the ROI of the chart line. The challenge is that the (0, 0) coordinate in the Cartesian plane is on the bottom left corner while the (0, 0) position in the image coordinate is on the top left corner, so the line coordinate must

be converted. After this process, all the texts from the chart is extracted and stored in a String variable and the pair (*x, y*) features are stored in an integer variable.

The next process is Natural Language Processing. The input for this process is the chart feature variable. The process is done by developing the sentence based on the seed words extracted from chart image. A Seed word is selected by the user to make sure that the system is only developing the sentence needed by the user. The generation is corpus based and the calculation is based on Bi-Gram probability. The result from this process, including the result from this overall system, is the sentence that describes the chart automatically.

1.5 Hypothesis

This thesis proposed a method for interpreting chart image automatically using the combination between image processing and Natural Language Processing (NLP). On image processing side, the OCR method proposed in [2][3][4] can be used to extract and recognize the character contained on the image. Then the OCR is used to extract the text from the chart such as chart title, axis title, axis number, and legend. Chart value is extracted using the calculation of chart line pixel according to the chart scale pixel position [5].

In [6], [7], [8], there are discussions about method to perform an automatic Natural Language Generation System. Based on their researches, the automatic chart interpretation system is realized using Natural Language Processing. The combination using Bi-Gram sentence and grammar rule methods results the more appropriate sentences because the sentences not only constructed using the sequence of sentences, but also calculates the grammar rule to make sure the grammar suitability.

The improvement from the previous methods are that the input can be in a form of image and the information which is used for the knowledge uses not only the corpus, but also information from the chart legends, title and values extraction, in addition to the corpus.

1.6 Assumption

According to the limitation of this thesis, then it is assumed that the chart image is not tilted and its quality is not degraded, for example, by noise. To make sure that the process can run properly, it is assumed that the original data is available and based on that data, the chart can be generated by the Matlab software. Information in the chart image is composed of chart title, legend, axis title, axis number, and chart line. It is assumed that the difference between axis, background, and chart line is clean and clear.

1.7 Scope and Delimitation

The chart is limited to two dimension (2D) chart, which is in the form of time series Chart (no two pair of (*x, y*) value which is exactly the same). Another limitation is that there is only one curve on the chart.

1.8 Importance of the Study

This study will contribute to automatic chart interpretation systems, which can help one in evaluating image-based data. The developed tool can also be regarded as an expert system or decision support system because it is able to produce recommendation based on its interpretation on the image data.