# Document Zoning for Enhancing Spatial and Temporal Understanding in Web-based Health Surveillance Systems

Hutchatai Chanlekha

DOCTOR OF
PHILOSOPHY

Department of Informatics,
School of Multidisciplinary Sciences,
The Graduate University for Advanced Studies (SOKENDAI)

2010

March 2010

A dissertation submitted to

The Department of Informatics,

School of Multidisciplinary Sciences,

The Graduate University for Advanced Studies (SOKENDIA)

In partial fulfillment of the requirements for

The degree of Doctor of Philosophy

Supervisor:

| | |
|---|---|
| Nigel Collier, *Assoc. Prof.* | National Institute of Informatics, SOKENDAI |

Advisory Committees:

| | |
|---|---|
| Nobuhiro Furuyama, *Assoc. Prof.* | National Institute of Informatics, SOKENDAI |
| Asanobu Kitamoto, *Assoc. Prof.* | National Institute of Informatics, SOKENDAI |
| Ken Satoh, *Prof.* | National Institute of Informatics, SOKENDAI |
| Hideaki Takeda, *Prof.* | National Institute of Informatics, SOKENDAI |
| Thanaruk Theeramunkong, *Assoc. Prof.* | Thammasat University (Thailand) |

# Abstract

Public concern over the spread of infectious diseases such as avian H5N1 influenza and swine flu (H1N1) influenza A has underscored the importance of health surveillance systems for the speedy and precise detection of disease outbreaks. However, two key barriers faced by the current web-based health surveillance systems are their inability to (a) understand complex geo-temporal attributes of events and (b) to obtain the levels of geo-temporal recognition. In this thesis, I develop a novel framework as a means to overcome these limitations. This framework is called spatiotemporal zoning.

The objective of the spatiotemporal zoning scheme is to enable language technology software to partition text into segments based on the spatiotemporal characteristics of its content. Each segment, which is called a text zone, contains a set of events that occurred at the same geographical location in the same time frame. The capability of associating events reported in each text segment with the most specific spatial and temporal information available in news reports enables simple techniques to be employed for detecting specific outbreak locations. These techniques could be, for example, text classification to detect text segments that indicate outbreak situations. At the same time, false alarms about past outbreaks can be avoided by taking the temporal information about the events into consideration.

I created a representative corpus in order to demonstrate that spatiotemporal zoning can be automatically and manually applied to unrestricted text. The corpus consisted of 100 news articles from multiple news agencies reporting on various disease outbreaks in different parts of the world.

To study the reliability of spatiotemporal zoning, an experiment was conducted in which three annotators were recruited to annotate the same set of documents according to the annotation guidelines and the agreement between these annotators was then analyzed. Several statistical measures, namely kappa, Krippendorff's alpha ($\alpha$), and the percentage agreement, were used for quantitatively measuring the agreement. The results showed that the level of agreement kappa was more than 0.9 on average for event type and temporal attribute annotations, and it was only a slight lower for annotating spatial attributes.

The task of spatiotemporal zoning can be separated into 3 main steps. (1) Document pre-processing: This step provides the basic elements for zone attribute analysis and was done automatically using natural language processing software. (2) Zone attribute annotation: Each event-predicate is analyzed to recognize its class, spatial and temporal attributes. (3) Zone boundary generation: This step is done based on the attribute values of each event-predicate. For spatiotemporal zone annotation, the study of automatic zone attribute annotation was done for each group of zone attributes, i.e., event type recognition, temporal attributes recognition, and spatial attribute recognition.

To automatically classify event expressions, i.e. zone type recognition, Conditional Random Fields (CRFs) was used to incorporate various sets of text features into a classifier.

To recognize spatial information, several approaches, ranging from simple techniques such as the commonly used heuristic-based approach to the more sophisticated machine learning approach, were experimented. I also explored various feature sets and feature encoding strategies in order to determine the best ones for recognizing spatial attributes.

For temporal attribute recognition, I took a rule-based approach to recognizing an event's temporal information. However, one of the problems is that in many cases the same event is repeatedly mentioned whereas the time of its occurrence is stated only once. To improve the system's ability to recognize the temporal information, I employed a simple heuristic that helps to identify linguistic expressions referring to the same events.

The above studies that I undertook prove that spatiotemporal zoning is reliable. Moreover, the results from automatic zone attribute recognition show that this scheme can be done automatically with a reliable level of performance.