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 (α), 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.

Acknowledgement

I would like to express my gratefulness to many people for their support, encouragement and guidance during my years as a graduate student of Sokendai at National Institute of Informatics.

First and foremost, this dissertation represents a great deal of time and effort not only on my part, but on the part of my supervisor, Nigel Collier, whose encouragement, guidance and support from the initial to the final level enabled me to shape my research as well as to develop an understanding of being a good researcher. This dissertation would not have been possible without him.

I also thanks to my other committee members, Ken Satoh, Hideaki Takeda, Asanobu Kitamoto, and Nobuhiro Furuyama for their valuable time discussing and making comments regarding my research.

I owe my deepest gratitude to all, the past and current researchers, in my research group. I would like to thank to Mike Conway, Doan Son and especially Ai Kawazoe for their support and comments

Many thanks to every people at NII, especially Hiroko Tokuda, Reiko Okano, Kumiko Ito, Miyuki Kobayashi and Yasuko Umebayashi who helped me carry on my research smoothly without irritating any office works. I am appreciated for your kindness and recommendations for solving my daily-life problems during my studies in Japan.

I am indebted to many of my colleagues in NII and Thailand to support me, cheering me up at all times when I needed them during my studies.

Most of all, I would never forget my mother has made available her support and encouraging in a number of ways through the entire period.

Contents

Abstract ..... **- 3 -**
Acknowledgement ..... **- 5 -**
Contents ..... **- 6 -**
List of Figures ..... **- 9 -**
List of Tables ..... **- 10 -**
Chapter 1 Introduction ..... **- 11 -**
1.1 Motivation ..... - 12 -
1.2 Objectives and Approach ..... - 15 -
1.2.1 Thesis Question ..... - 15 -
1.2.2 Approach ..... - 15 -
1.3 Contributions ..... - 15 -
1.4 Reader's Guide to the Thesis ..... - 16 -
Chapter 2 Background and Related Work ..... **- 17 -**
2.1 Previous works on spatiotemporal analysis in text ..... - 17 -
2.1.1 Temporal analysis in text ..... - 17 -
2.1.2 Spatial analysis in text ..... - 20 -
2.1.3 Spatiotemporal analysis in text ..... - 22 -
2.2 Previous works on text content analysis ..... - 24 -
2.2.1 Text Tilling ..... - 24 -
2.2.2 Argumentative Zoning ..... - 25 -
2.3 Basic idea for Spatiotemporal zoning design ..... - 25 -
2.4 Background theories and studies for Spatiotemporal zoning design ..... - 28 -
2.4.1 Linguistic component conveying temporal information ..... - 28 -
2.4.2 Temporal reference point and its interpretation ..... - 29 -
2.4.3 Temporal move of reference point ..... - 30 -
2.5 Characteristics of news report on disease outbreaks ..... - 31 -
Chapter 3 Spatiotemporal Zoning ..... **- 34 -**
3.1 Definition of events ..... - 34 -
3.2 Spatiotemporal Zoning: Task definition ..... - 35 -
3.3 Zone attribute: Class of news content ..... - 36 -
3.3.1 Generic information ..... - 37 -
3.3.2 Hypothetical event ..... - 38 -

3.3.3 Temporally-locatable event ..... - 38 -
3.4 Issues in Spatiotemporal zoning framework design ..... - 40 -
3.4.1 Issue in news content classification ..... - 40 -
3.4.2 Issue about temporal attribute ..... - 40 -
3.4.3 Temporal granularity in health news events ..... - 42 -
3.4.4 Spatial granularity in health news events ..... - 42 -
3.5 Spatiotemporal zoning annotation scheme ..... - 42 -
3.5.1 Zone attributes ..... - 44 -
3.5.2 Metadata ..... - 48 -
3.5.3 Unit of annotation ..... - 49 -
3.5.4 Nested annotation ..... - 51 -
3.5.5 New zone instantiation ..... - 52 -
Chapter 4 Evaluation of the Spatiotemporal Scheme ..... - 53 -
4.1 Data set ..... - 54 -
4.2 Annotation setting ..... - 57 -
4.2.1 Annotators ..... - 57 -
4.2.2 Annotation tool ..... - 57 -
4.2.3 Annotation guideline design process ..... - 58 -
4.2.4 Process of annotation experimentation ..... - 58 -
4.3 Agreement measurements ..... - 59 -
4.3.1 Kappa ..... - 59 -
4.3.2 Percentage agreement ..... - 60 -
4.3.3 Krippendorff's alpha ..... - 60 -
4.4 Reliability studies ..... - 61 -
4.4.1 Qualitative analysis of each zone class ..... - 61 -
4.4.2 Quantitative analysis of zone class annotation ..... - 65 -
4.4.3 Zone spatial attribute annotation results ..... - 70 -
4.4.4 Zone temporal attribute annotation results ..... - 74 -
Chapter 5 Automatic Approaches for Spatiotemporal Zoning Annotation ..... - 77 -
5.1 Zone generation process ..... - 77 -
5.2 Automatic approach for event type annotation ..... - 79 -
5.2.1 Previous approach for text content classification ..... - 79 -
5.2.2 Event type annotation approach ..... - 80 -
5.2.2.1 Conditional Random Fields ..... - 80 -
5.2.2.2 Information source for event classification ..... - 81 -

5.2.3 Event classification Results ..... - 83 -
5.3 Automatic approach for spatial attribute annotation ..... - 86 -
5.3.1 Information source for spatial attribute recognition ..... - 87 -
5.3.2 Spatial attribute recognition: Experimentation ..... - 90 -
5.3.3 Discussion ..... - 103 -
5.4 Automatic approach for temporal attribute annotation ..... - 104 -
5.4.1 Temporal attribute recognition approach ..... - 105 -
5.4.2 Temporal attribute recognition: Experimentation ..... - 107 -
Chapter 6 Conclusion in Future Work ..... - 109 -
Future studies ..... - 110 -
About Author ..... - 112 -
Related Publication ..... - 113 -
Bibliography ..... - 114 -

List of Figures

Figure 1-1: Various locations with different roles in outbreak news reports ..... - 13 -
Figure 3-1: Text capture of spatiotemporal zoning in a news report ..... - 36 -
Figure 3-1: Temporal metadata in the spatiotemporal zoning scheme ..... - 49 -
Figure 4-1: Distribution of the number of sentences, including partial sentences ..... - 55 -
Figure 4-2: Distribution of the number of events to be annotated ..... - 55 -
Figure 4-3: Distribution of news articles in the corpus in terms of the publication date ... - 56 -
Figure 4-4: Distribution of outbreak situations reported in the corpus, classified in terms of outbreak-affected country ..... - 56 -
Figure 4-5: Interface of the spatiotemporal zone annotation tool ..... - 57 -
Figure 4-6: Example of co-referring of events ..... - 75 -
Figure 5-1: Zone generation process ..... - 78 -
Figure 5-2: Algorithm for calculating feature score ..... - 93 -
Figure 5-3: Spatial attribute recognition results from the models trained with different combinations of spatial-related textual features ..... - 100 -

List of Tables

Table 3-1: Zone attribute for spatiotemporal annotation ..... - 43 -
Table 3-2: Constituent types qualifying for an independent zone annotation ..... - 51 -
Table 4-1: Data statistics ..... - 54 -
Table 4-2: Proportions of events classified by each annotator ..... - 65 -
Table 4-3: Krippendorff's diagnostics for category distinction ..... - 66 -
Table 4-4: Confusion matrix between annotators A and B on Set1 and between annotators A and C on Set2 ..... - 66 -
Table 4-5: Agreement statistics (percent agreement) for location attribute annotation .... - 70 -
Table 4-6: Agreement statistics for temporal attributes annotation ..... - 74 -
Table 5-1: Linguistic features for event classification ..... - 81 -
Table 5-2: Evaluation of the contribution of contextual features in event classification . - 83 -
Table 5-3: Evaluation of the contribution of internal features in event classification ..... - 84 -
Table 5-4: Proportion of events that were miss-classified by the model ..... - 85 -
Table 5-4: Experimentation results for recognizing spatial attribute of the events based on the heuristic approach ..... - 91 -
Table 5-5: Experimentation results for recognizing spatial attribute of the events based on the probabilistic approach ..... - 93 -
Table 5-6: Feature encoding method for the learning task ..... - 95 -
Table 5-7: Experimentation results for recognizing spatial attribute of the events based on the statistical machine learning approach ..... - 96 -
Table 5-8: Experimentation results of spatial attribute recognition from the models trained with different combinations of spatial-related textual features ..... - 98 -
Table 5-9: Examples of rules for recognizing event's temporal attribute ..... - 106 -
Table 5-10: Experimentation results of temporal attribute annotation ..... - 107 -

Chapter 1

Introduction

The International Health Regulations (2005) [1], which entered into force on 15 June 2007, have bound 194 countries around the globe to a new legal framework for coordination of the management of events that may constitute a public health emergency of international concern. The implementation of the framework has underlined the importance of health surveillance technology, both indicator-based, using structured data collected through routine health surveillance, and event-based, using unstructured text sources. Despite advances in indicator-based public health surveillance [2, 3], under-developed public health systems are a significant barrier to compliance in many parts of the world [4-6].

Event-based surveillance systems have become another crucial source of epidemic surveillance to fill this gap. Examples of event-based systems including MedISys [7] (EU), GPHIN (Canada) [8, 9], Argus [10], EpiSpider [11], HealthMap [5], and BioCaster [12, 13]. These systems look for outbreak signals in a variety of electronic sources, including news wire, official reports, and email, which can provide localized and near real-time data on disease outbreaks [4, 14, 15]. The unstructured texts that are found are then processed using automatic text mining. Any outbreak-related information that is extracted is organized and presented to users. Most systems provide map-based visualizations by geo-coding alerts to the country scale, with province-, state-, or city-level resolutions for selected countries [5, 7, 11-13, 16].

The need for better surveillance became more urgent during the past decade with the recognition of the potential influenza pandemic and also the threat of accidental or deliberate release of hazardous agents such as anthrax. However, one of the limitations that the current event-based health surveillance systems face is their inability to improve the performance regarding to the spatial granularity of automatically detected outbreaks. The ability to identify the location of an outbreak at the finest level of granularity is very crucial in various aspects. For example, detail information about the outbreak's location can help analysts to monitor and study the evolution of the outbreak. Identifying an

outbreak at a coarse spatial granularity potentially hides important developments in the spread of an event to neighboring towns within an area. In the example below, it is far more useful to know that the outbreak occurred in a city or town than in a country.

> Avian influenza was confirmed in Viet Nam
> The Ministry of Health in Viet Nam has provided WHO with official confirmation of an additional eight human cases of H5N1 avian influenza.
> Two of the cases were recently detected, between 2 and 8 April, in Hung Yen and Ha Tay Provinces, respectively.

In the above example, most systems that employ a naïve strategy for automatic outbreak detection would recognize Viet Nam as a site of the avian influenza outbreak. However, it would be more beneficial for monitoring and situation analyzing if a system could recognize the province where the outbreak occurred.

This thesis addresses this problem by focusing on public health surveillance. It describes a new scheme with the purpose to overcome the current limitations on enhancing not only the spatial information, but also temporal information of events.

1.1 Motivation

Geo-temporal encoding of outbreak reports at fine levels of granularity is one of the key requisites for greater utilization of event-based health surveillance systems, but can now only be achieved with accuracy by hand encoding of reports which is time consuming and expensive. For automatic encoding, current systems tend to adopt ad-hoc strategies, generally in the form of detecting the first disease and location pair that match predefined criteria or similar heuristics in order to identify the disease-affected location, and use publication date as an approximate occurrence time of the outbreak events. Although these strategies are effective in reducing both computational time and false alarming of outbreaks in irrelevant locations, they may lead to under-reporting of events or issuing reports at sub-optimal levels of granularity. This results from a characteristic of news, in which the details of the specific locations, as well as the exact time of an outbreak are usually mentioned later in a story.

To improve the performance of event-based health surveillance systems, we need to go beyond heuristic methods that analyze only headlines or the first few sentences of documents. It has been reported, however, that blindly searching for locations in full text,

while increasing the detection sensitivity, can lead to excessive false positives [16]. This is because a news story about an outbreak may also refer to locations that are not directly affected but are related to the situation in various ways, e.g., neighboring countries that might be impacted, countries that provide medical assistance, previously affected locations, and so forth. The text capture shown in figure 1-1 exemplifies this situation. In figure 1-1, the outbreaks were reported to occur in *Nunavut* (in example1) and *Botswana* (in example2). Other appearances of location names, such as *Japan*, *Caribbean countries*, and *Africa* (in example1) are referred to as locations where HTLV-1 usually occurs, while *South Africa* and the *U.S.* (in example2) are countries providing medical assistance to the affected country.

EXAMPLE 1

...
Health officials in Nunavut will begin asking all expecting and nursing mothers to undergo screening for a rare and untreatable virus that has appeared in the territory.

Human T-cell Lymphotropic Virus, Type 1 (HTLV-1) occurs mostly in Japan, Caribbean countries and Africa, affecting between 15 million and 25 million people worldwide.

Doctors say most people who contract it will show no symptoms, but in about five per cent of cases, it can lead to cancers of the blood and diseases affecting the nervous system. The development of those conditions can take 10 to 20 years.

It's rarely seen in Canada, but as many as 20 people in Nunavut have tested positive for it.
...
...

EXAMPLE 2

...
Since January 2006 diarrhea in Botswana has claimed the lives of 446 babies.

Botswana is struggling to contain the disease and is currently getting assistance and expertise from South Africa, the Centers for Disease Control in the U.S., the World Health Organization and other international health organizations.
Over 11,000 children have already been affected by the disease.
...

Figure 1-1: Various locations with different roles in outbreak news reports

To identify outbreak locations with finer granularity while minimizing false alarms as much as possible, it is necessary to use a more sophisticated approach that enables systems to distinguish "locations" where the "current" outbreak is occurring from other locations. More specifically, the framework must as a minimum provide means to (1) identify outbreak locations at the finest level of granularity offered by the text and (2)

distinguish newly reported data from historical and hypothetical data. Recognizing the precise time of the outbreak faces a similar issue: Although there may be a lot of temporal expressions mentioned in the documents, some of them might not be directly related to the outbreak situation.

One existing linguistic-oriented approach that is capable of overcoming these limitations is information extraction [17-19], which analyzes documents and extracts outbreak-relevant information, such as the disease, location, and time. However, the inherent problem that any information extraction systems generally face is a trade-off between specificity and sensitivity. Since the precisions of outbreak detection is very important in health surveillance systems, information extraction employed in such systems tends to have high specificity, which generally leads to a failure in detecting a number of outbreak affected locations. For example, the sensitivity of one reported information extraction system for outbreak reporting was less than 50% [17].

To overcome the limitation of the existing systems while at the same time alleviating the problem of information extraction, I hypothesize that the time and place of an outbreak can be extracted through the capability to associate events reported in each text segment with the most specific geographical and temporal information available in news reports. Through this association, outbreak-affected locations can be identified by using simple techniques such as 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 events into consideration. This thesis proposes a novel framework called spatiotemporal zoning. Spatiotemporal zoning integrates analyses of spatial and temporal attributes of events in order to mitigate the inherent limitations of current surveillance systems.

Despite their tendency for rumor mongering, the evidence that news reports enabling real-time outbreak detection has underscored its necessity as one of important sources for health surveillance. For example, it has been recorded that during 1994 epidemic of pneumonic plague in India, the epidemic reports on CNN International were timelier at the initial stages of the outbreak than official public health sources elsewhere [20]. For that reason, this thesis focuses on the development of a framework for analyzing news reports.

1.2 Objectives and Approach

1.2.1 Thesis Question

The principal question addressed in this thesis is:

For the task of document-based health surveillance, how can spatiotemporal information be represented and acquired automatically?

1.2.2 Approach

To solve this critical issue, I have developed a novel approach that is capable of analyzing text according to its spatiotemporal characteristics. The studies and techniques of linguistics and statistical machine learning were employed in this development. More specifically, two main subsidiary works have been done.

- **Construction of the spatiotemporal zoning schema, including its specification, annotation guidelines, and gold standard corpus**: Spatiotemporal zoning was designed based especially on theories relating to events and time, and the target application requirement. To evaluate the scheme and the automatic annotation approaches that could be used with it, it is necessary to have a gold standard corpus. A gold standard corpus, an annotation specification, as well as an annotation tool were thus developed for this purpose.
- **Development of automatic system for spatiotemporal zoning**: The ultimate purpose of this thesis is to deploy the spatiotemporal zoning framework in operating online health surveillance systems. For this to be possible, spatiotemporal zoning will have to be done automatically. Various automatic approaches were experimented with, ranging from simple rules-based and probabilistic approaches to sophisticated machine learning approaches.

1.3 Contributions

This thesis makes four distinct contributions to the fields of linguistic and information processing.

- **Spatiotemporal zoning framework**: This novel framework combines analyses of the spatial and temporal information contained in text. The thesis proves

framework is reproducible and that an automatic annotation system embodying this framework can be developed with promising performance.

- **Resources for spatiotemporal zoning**: The resources developed in this thesis consist of a gold standard corpus of 100 spatiotemporal zone annotated outbreak news reports. Spatiotemporal zone annotation guidelines and a zone annotation tool were also created for the gold standard corpus development task.
- **Automatic approaches to zone annotation**: The task of zone annotation was separated into 3 main subtasks for recognizing each zone attribute, i.e. zone class, spatial attribute, and temporal attribute. This thesis explores various methodologies for automatically annotating each spatiotemporal zone attribute. It also discusses and evaluates the textual features to be used in the recognition model.

1.4 Reader's Guide to the Thesis

The remaining chapters of this thesis are organized as follows.

- **Chapter 2** provides background knowledge on the design of the spatiotemporal zoning framework. The related studies are discussed.
- **Chapter 3** describes the spatiotemporal zoning framework. The issues faced in the design as well as the annotation specification are also discussed.
- **Chapter 4** describes how the reproducibility of the framework was assessed. The experimental setting and annotation results are presented. A thorough analysis of the results is given.
- **Chapter 5** discusses automatic approaches for spatiotemporal zone annotation. Automatic approaches for each zone attribute are described and evaluated. Also illustrated is a strategy for the results of the zone annotation to be used to enhance the spatial granularity of information in a health surveillance system.
- **Chapter 6** presents conclusions and outlines future work.

Chapter 2

Background and Related Work

Before introducing the propose scheme, in this chapter, I first discuss the background knowledge and published work related to spatiotemporal zoning. The development of this scheme benefitted from various ideas in linguistics, such as the notion of events, temporal interpretation, and the relation of events and time. Previous studies on text content analysis that influenced the design of this work are also discussed in this chapter.

This chapter is organized as follows. First, the previous studies on spatiotemporal analysis are investigated. Then, the basic idea of spatiotemporal zoning is introduced. After that, the related work and linguistic ideas that strongly influence the scheme design are discussed. Finally, the characteristics of news reports, especially disease outbreak reports, are described.

2.1 Previous works on spatiotemporal analysis in text

There are many works that proposed means to analyze spatial and temporal information in text content. Generally, these works provided the analysis of spatial and temporal information separately. Only a few works that combined spatial and temporal together as a unify framework. In this section, the previous works, which focused on provide understanding of spatial and temporal information in text, are discussed.

2.1.1 Temporal analysis in text

Texts, such as newspaper, narratives, etc, almost always describe events which occur in time and specify the temporal location and order of these events. Text comprehension involves the capability to identify the events described in text and locate these events along timeline. Temporal understanding of text is very important in perceiving occurring sequence and relation of the events expressed in text. Temporal interpretation gained a lot of attention in various fields, e.g. linguistic, psycholinguistic, artificial intelligence. There were many works that proposed theory about events and time, as well as temporal

characteristics and interpretation in text [21-31]. In computational linguistic field, the introduction of TimeML [32] has laid a robust framework and brought a lot of attention from many research groups for automatic analyzing event and time in natural language text.

Temporal analysis framework: TimeML

TimeML [32] is a rich specification language for event and temporal expressions in natural language text. Through several rounds of specification revisions, it can now be considered a gold standard for temporal information. TimeML captures three phenomena in temporal markup: (1) It systematically anchors event predicates to a broad range of temporally denoting expressions. (2) It orders event expressions in text relative to one another. (3) It allows for a delayed (underspecified) interpretation of partially determined temporal expressions.

The task of temporal analysis in TimeML concerned four aspects in event and temporal expression markup, which are (1) identifying an event and anchoring it in time; (2) Ordering events with respect to one another; (3) Reasoning with contextually underspecified temporal expressions; (4) Reasoning about the persistence of events. Various works have been proposed to solve different problems of temporal annotation.

Although TimeML framework provides very rich and useful information regarding to the temporal side of event, it is quite complex and considered to be impractical to be able to develop a reliable automatic system that is capable to do the whole task [33]. The temporal aspect of the spatiotemporal zoning framework was strongly influenced by TimeML. However, I have tailored the proposed framework to make it simpler, which is confirmed by the improvement of inter-annotator agreement scores. Despite the simplicity, I try to preserve enough temporal information to facilitate the rapid distinguishing newly reported information from previously reported data.

Automatic temporal analysis

Evita [34] is the system that proposed to solve the one problem within TimeML, which is identifying an event expression in text. The functionality of Evita breaks down into two parts: event identification and analysis of the event-based grammatical features that are relevant for temporal reasoning purposes, such as tense, aspect, polarity, and modality. TimeML identifies as events those event-denoting expressions that participate in the narrative of a given document and which can be temporally ordered. The

identification of event that represented by verbal chunks is based on lexical look-up, accompanied by minimal contextual parsing in order to exclude weak stative predicates, such as 'be', and some generics (e.g., verbs with bare plural subjects). Identifying events expressed by nouns involves two parts: a phase of lexical look-up; and a word-sense disambiguation process based on statistical approach. Identification of events from adjectives takes a conservative approach of tagging as events only those adjectives that were annotated as such in Time-Bank1.2, whenever they appear as the head of a predicative complement.

The work in [35], focus on the problem of learning temporal relations between event and event; and event and time. In this work, they considered temporal relation, i.e. TLINK, labeling as the following classification problem: given an ordered pair of elements X and Y, where X and Y are events or times which the human has related temporally via a TLINK, the classifier has to assign a label of relation type or none to the link. They employed machine learning techniques for TLINK classification, which are Maximum Entropy and Conditional Random Fields. The works in [36] also concerned the same problem of temporal relations recognition. In this work, they employed Maximum Entropy for feature learning, as well as tried on other classifiers, such as naïve Bayes, Support Vector Machine. The features used in this work were, for each event in an event-ordering pair, the *event-class*, *aspect*, *modality*, *tense and negation*; *event string*, and *signal* (a preposition/adverb, e.g., reported on Tuesday), which are string features. There were also contextual features indicating whether both elements in the event pair have the same tense and same aspect. For event-time links, they encoded the above event and signal features along with TIMEX3 time features.

In TempEval track, which is an evaluation exercise for temporal-related annotation under SemEval-2007 [33], the task of TimeML was reduced into three subtasks, which are recognizing:

Task A: temporal relation holding between event expressions and time that occur within the same sentence

Task B: temporal relation holding between document date and event expressions

Task C: temporal relation holding between main events of adjacent sentences

In these tasks, inter-annotator agreements[1] were quite low, 0.72 for task A and B, and 0.65 for task C.

TempEval task provided a simple standard to evaluate automatic extraction of temporal relations. In this evaluation, there were six teams [37-42] that participated to solve different aspects in temporal relation annotation task. The approaches employed by these participants were rule-based approach, statistical/machine learning approach, and a hybrid between rule and statistical approach.

2.1.2 Spatial analysis in text

Location expression recognition

The recognition of location expressions in text was firmly defined in Message Understanding Conference as a part of the task called named entity recognition (NER). In NER, linguistic expressions in text are recognized and classified into predefined types of entities, such as person, organization, location, date and time, etc. The introduction of NER has brought a lot of attention from various research groups in exploring automatic approaches for recognizing named entity in text [43-48].

The approaches for named entity recognition can be categorized into 3 groups, which are hand-crafted rule-based approach, statistical and machine learning approach, and hybrid approach.

NER systems that employed the hand-crafted rule-based approach usually based on one of the three methodologies [49], which are:

1) Recognizing named entities by using regular expression rules that were constructed by human experts [50, 51].

2) Utilizing external knowledge sources, such as name lists, gazetteers, dictionary, for identifying named entity. The recognition of named entity is done by searching for matches between expressions in text and name expressions in dictionary [52].

3) Using sophisticated linguistic rules [53, 54] or heuristic rules [55].

The second class of methodology is to develop a statistical model for recognizing named entities in text. Automatic learning approaches for NER have been proposed in

[1] Agreement was measured by the average of precision and recall

order to reduce time and effort of human expert in developing the systems or porting systems to new domains or languages. The machine learning approach learns knowledge for recognizing named entities from the real characteristics of NE in training texts.

Various machine learning techniques that have been employed in named entity recognition task. These techniques are, such as Hidden Markov Model (HMM) [56], Maximum Entropy [57-59], Decision tree [49, 60, 61], Support Vector Machine (SVM) [62, 63]. The machine learning techniques widely used for NER are usually in the form of supervised learning, which needs pre-annotated training materials to be used as answer keys. However, there were many works that concerned about the disadvantage of supervised learning when there is no training corpus to use in the learning task. To tackle this limitation, many works proposed methodologies that employed minimally-supervised or unsupervised techniques in order to recognized named entity [64, 65]. Basic idea of these approaches is based on the bootstrapping technique that utilized the redundancy of named entity occurring patterns.

The hybrid approach for NER is a couple between hand-crafted rule-based approach and statistical machine learning approach. The example of the system that employed this approach is LTG [66], which used rules together with statistical partial matching technique.

Geographical grounding

One of the limitations in NER task is its ignorance in relating the text span to a model of the world. However, the grounding of spatial attribute of events is a precondition for accurate understanding and reasoning. It provide critical information in solving location ambiguity and precisely visualizing geographical data Hence, many researches were proposed to solve this problem by attempting to associate textual location expression with the real world geographical location, which can be, such as geographical ontology, or geographic coordinates [67-71]. The geographical grounding approaches usually involve geographical gazetteers and based on the one-sense-per-discourse assumption.

Many strategies were proposed to disambiguate location name appeared in text. For example, in [68], heuristics were used for solving ambiguous place names. They applied two different *minimality heuristics*. The first one was adapted from works in automatic word sense disambiguation, which is one-sense-per-discourse assumption. The other one was the assumption that, in cases where there is more than one place name mentioned in

some span of text, the smallest region that is able to ground the whole set is the one that gives them their interpretation. In [70], hybrid approach was used for normalizing place names. The hybrid approach combined (i) lexical grammar driven by local context constraints, (ii) graph search for maximum spanning tree in order to choose the best matching sense set and (iii) integration of semi-automatically derived default senses. In [72], co-occurrence model was used for location normalization. This work built co-occurrence model of how place names occur together in Wikipedia then applies the co-occurrence model to disambiguate the named entities.

Although these works provide critical information in solving location ambiguity and precisely visualizing geographical data, they do not associate these locations to events or actions mentioned in text. Nevertheless, the ability to anchor events with the locations in which they occurred would is important for an effective analysis and understanding of the situations being reported in articles.

2.1.3 Spatiotemporal analysis in text

Although the works in analyzing spatial and temporal information of text are usually separated from each other, there are also several works that combined the analysis of the two event attributes together.

The group of works that, although do not fully analysis of time and place of every event in text, are capable of recognizing place and time of certain set of predefined events is information extraction (IE). Information extraction is the task that analyses unrestricted text in order to automatically extract structured-information about pre-specified types of events, profile of entities, or relationships from a certain domain. The works that extract information about certain events from news reports usually encode spatial and temporal of the events in the extraction template. For example, in the template defined in MUC-7 [73] for extracting information about vehicle launching event, there are “LAUNCH_DATE” and “LAUNCH_SITE” slots for identifying time and place of the extracted launching events. In [17, 18], which focus on extracting information in outbreak domain, also encoded, among other outbreak-related information, information about place and time of the outbreak in the extraction template. To extraction information, various approaches have been proposed by different research groups. The works on information extraction usually rely on rule or pattern-based approach, either automatically constructed or manually created. However, there are also works that

used machine learning model, such as Hidden Markov Model [74], Support Vector Machine [75], for extracting relevant information.

Rules used in information extraction systems can be constructed manually or automatically. Both supervised and unsupervised strategies were explored for automatic rules construction. In [76-78], rules were constructed from pre-annotated corpus. Another group of works created extraction rules from unannotated text, with some initial domain-specific knowledge, i.e. keywords, or initial handcrafted rules) and/or some validations from the user in order to learn effectively [79-83]. In [82, 84-88], bootstrapping method was used in order to extract entities in the category of interest to be filled in the IE template slots. Many works employed semi-supervised approach in order to extract specific relations between entities [89, 90]. These studies focus on extracting relations between pairs of instances of particular categories. The task usually starts by searching for sentences that contain both entities' instances and then learns the patterns associated with the two entities.

Although the IE works are capable of recognizing spatial and temporal attribute of events, but they focused only on a predefined set of events. Because of the awareness of the difficulties for question-answering (QA) systems or decision making systems in analyzing and processing complex events in text, some works proposed a formal representation of events, which combined together the basis event attributes, namely space and time. Chaudet [91] extended the event calculus and proposed the SpatioTemporal Extended Event Language (STEEL) that is based on joint spatial and temporal location of event occurrences. STEEL is a typed first-order logic language. The objective of the proposed language is to mediate the representation of textual content from strict information extraction templates and free document semantic representation with flat logical forms. It provides a logical representation of report contents in order to express the outbreak history at the semantic level for building a qualitative model of the epidemics. It also allows the representing and building event aggregates according to the spatiotemporal location of their occurrence. This specification language provides 10 predicates for modeling the spatial and temporal relationships between spatiotemporal positions. These predicative relations are; "disconnected", "part of", "proper part of", "identical with", "overlaps", "discrete from", "partially overlap", "externally connected to", "tangential proper part of", "non tangential proper part of". STEEL expressed a deep insight into the understanding of events in text. However, the representation according to

this work is sophisticated. Hence it is quite difficult to develop an automatic system that can process free text into the STEEL representation. At the time of writing this thesis, there is no evidence of practical implementation of this language. The work presented in this thesis could be thought of as a simplified version of this formal representation. The results from spatiotemporal zoning provide, at certain level, basic information about events which can be further elaborated into the STEEL representation.

Schuurman [92] proposed a framework that introduced spatiotemporal layer of annotation to be added to Treebank corpus. This work attempted to locate eventualities on a time-axis and to disambiguate geospatial information such that the event entities can be located on a map. Basically, while this work located events on a time-axis, it did not concern about associating event with its spatial attribute. The processing of spatial information was in the form of geographical grounding.

2.2 Previous works on text content analysis

This section described about the previous works that are related to text content analysis and text content classification. Brief introduction of these works, as well as the coverage of their frameworks to the problems focused in this thesis was also discussed.

2.2.1 Text Tilling

Text-tilling [93, 94] attempts to discover coherent, interrelated sub-discussions to reflect the pattern of subtopics contained text. Basically, the task of text tilling is to partition text into coherent multi-paragraph units that discussing the same topic. It can be viewed as the task of organizing text content based on the topic of discussion. However, it does not provide information about what type of topic is discussed in each text portion, or when and where the discussed events happened.

This work resembles text tilling in that it also try to partition text. However, spatiotemporal zoning aims at portioning text into multiple units based on the type of text content, as well as spatial and temporal attributes of events expressed in text, regardless of discussed topic.

2.2.2 Argumentative Zoning

The purpose of argumentative zoning or rhetorical zoning is to analyze the global rhetorical status of each sentence or other constituent in the text and divide text into zones. Argumentative zoning was originally proposed with the aim of partitions text into rhetorical zones in terms of argumentation and intellectual attribution in a flat structure [95, 96]. They focus on computer science articles and classified text content into 7 classes, which are "aim", "background", "own" (i.e. detail of the solution), "contrast with other approaches/weakness of other approaches", and "basis" (i.e. imported solution). Their work was developed with the main purpose to facilitate the summarization of scientific articles. So it is important for the summarization system to recognize the intellectual attribution property of each text segment in order to effectively selecting the text portions to appear in the summary. The work of [97] also proposed a framework for argumentative zoning, but their framework partitions text into shallow nested structure based on rhetorical status, and focus specifically on scientific text in biological domain. Their proposed text categories are, "background", "problem setting", "method", "result", "insight", "implication", "else", "connection", "different", and "outline". The main purpose of this work is to facilitate information extraction systems for pinpointing and organizing factual information reported in biological articles.

Argumentative zoning provides an insight into what type of information can be expected in each zone. However, the characteristics and the nature of scientific text and news report article, which is one of the main informal sources for outbreak detection in major online surveillance systems, such as GPHIN, HealthMap, are not the same. The specifications of these works may not be suitable for directly apply to news articles. Spatiotemporal zoning was designed based on the idea of argumentation zoning; however, in stead of classifying text according to the argumentation attribution of the content, it classifies text according to the spatial and temporal attribute of the content.

2.3 Basic idea for Spatiotemporal zoning design

As mentioned earlier, the hypothesis that motivated this thesis is that the time and place of an outbreak can be extracted through the capability to associate events reported in each text segment with the most specific geographical and temporal information available in news reports. In order to effectively identify outbreak locations with better

granularity, while minimizing false alarms as much as possible, it is necessary to enable systems to distinguish "locations" where the "current" outbreak is occurring from other locations. More specifically, the framework must as a minimum provide means to (1) identify outbreak locations at the finest level of granularity offered by the text and (2) distinguish newly reported data from historical and hypothetical data.

As I looked deeper into this kind of data, I found that certain types of news content cannot be anchored along a timeline, and hence, cannot be associated with temporal information. Among these types of content are sentences that provide general knowledge about certain subjects, such as diseases, and sentences that predict or express the possibility of certain situations. Moreover, the content of these kinds of sentences also possesses a different informative level with respect to outbreak situation analysis. Given this observation, I believe that the capability to classify news content is necessary not only for event time identification but also as a means to enable more flexible content processing.

To deal with all the issues mentioned above, this thesis proposes a novel framework called spatiotemporal zoning, which integrates the classification of news content and analyses of the spatial and temporal attributes of events, as another means to mitigate the inherent limitations of current surveillance systems. In the designing, the framework has been tailored to satisfy the basic requirements of analyzing spatial and temporal information about events. At the same time, since the detection of outbreak alerts has a high impact on society and the economy, I have tried to keep the framework simple enough to allow for a reliable automatic system to be implemented, using extant knowledge sources such as part of speech taggers, named entity recognizer, shallow parser. The proposed framework tries to overcome the limitations of current health surveillance systems in the following ways:

1) Classifying news content into predefined content classes based on its spatial and temporal characteristic

As mentioned earlier, certain classes of news content cannot be located in time. To be able to effectively analyze the temporal attributes of reported events, it is necessary to provide a means to classify news content according to spatial and temporal characteristics. This task is similar to what is proposed in existing works on argumentative zoning [95, 97]. These works, however, focus on scientific articles and

classify texts into rhetorical zones in terms of argumentative and intellectual attribution. Since spatiotemporal zoning task focuses on classifying news report content in terms of its spatial and temporal attributes, direct application of these frameworks may not be suitable.

2) Recognizing the spatial attributes of an event

Spatiotemporal scheme associates each reported event with its location. Although the scheme does not directly extract outbreak locations, its capability of associating reported events with their locations allows for further processing to distinguish outbreak locations from other locations in a news story. Since all event-denoting linguistic expressions in text are associated with the name of location where the event occurred, an outbreak-affected location can be identified by detecting certain sets of words or expressions that are related to the outbreak situation. Automatic recognition of geo-temporal information about events in natural language texts, however, is not a trivial problem. For event location recognition, the preliminary study on 100 randomly selected news documents showed that a location of an event is the one referred by the location name closest to the event expression in the story only about 52% of the time. In order to correctly identify the locations of events, a more sophisticated technique applying syntactical analysis of sentences is needed.

3) Recognizing the temporal attribute of an event

In addition to spatial information, according to the proposed scheme, each event reported in text is also anchored to the approximate time that it occurred. Recognition of this temporal attribute is crucial in reducing false alarms of past outbreaks, since it provides information that systems require to distinguish newly updated information from previously reported information. For the temporal aspect of the proposed framework, I based my design on one of the existing frameworks for temporal processing, namely, TimeML [32, 98]. TimeML is a rich specification language for event and temporal expressions in natural language text and can be considered as a gold standard for temporal information processing. Although the TimeML framework provides very rich, useful information regarding the temporal aspects of an event, it is quite complex, and it is considered impractical to develop an automatic system that is capable of performing the whole task [33]. Therefore adopting its full expressivity may not be necessary for the spatiotemporal zoning task.

2.4 Background theories and studies for Spatiotemporal zoning design

There are a lot works in the area of linguistic, psycholinguistic, and cognitive science that concern about the events and temporal information represented in text. In this section, the studies and theories regarding to the relation between events and time as well as the temporal interpretation of text are introduced. This section introduces various kinds of basic knowledge and theories on which spatiotemporal zoning are based.

2.4.1 Linguistic component conveying temporal information

In narrative, without temporal adverbial or temporal conjunction, events conveyed in a sequence of sentences are usually interpreted as happening in the same time, happening in succession, overlapping with each other, or properly included in one another.

Partee [30, 99] introduced the theory of temporal anaphora which related to three types of linguistic components. These three are tense morpheme, temporal adverbial, and temporal conjunction.

- Tense morpheme

Tense morpheme is grammatical unit that conveys information about tense. For example: "One of the two new cases ***occurred*** in Oromia Province"

- Temporal adverbials

Temporal adverbials are phrases that explain when something happens. Smith [31] proposed a subcategorization of frame adverbial phrases based on the way in which frame adverbials depend for their semantic interpretation on the linguistic context of a discourse. Such context-dependency is called as "capture" in her work. There are three types of adverbials:

1) Adverbials which are protected from capture-adverbials in this group do not depend semantically on the linguistic context. Examples of adverbial in this group are, such as, last week, a week ago, yesterday, now, tomorrow, in three days, etc.

There is also another group of adverbials which are protected from capture. These adverbials can be called as complete or independent dates [26], such as, "in 1875", "in January 1990", "On April 4th, 1950", etc.

2) Dependent adverbials, which demand capture. Examples of adverbial in this group are, such as, previously, earlier, the same time, later, afterwards, etc.

3) Flexible anchoring adverbials, which are available for capture. Examples of adverbials in this group are, such as, on Tuesday, after Mike arrived, before James left, etc.

- Temporal conjunction

Temporal conjunction is sub-ordinate conjunction that is used in regard to time. Temporal conjunctions are, such as; after, before, when, while, since, until. Temporal conjunctive clause that many researches related to temporal-based discourse analysis are interested in is *when*-clause.

2.4.2 Temporal reference point and its interpretation

For temporal analysis, it is important to understand the concept of reference time in order to effectively interpret temporal information of the story. This thesis follows the terminology of Dowty [25]; where the "reference time" or "temporal reference point" is the time at which the event or state mentioned by the sentence occurred (or obtains, in the case of state). The time at which the sentence is heard or read by the hearer will be referred to as the speech time.

According to Hinrichs [26], the characteristics of temporal reference point can be summarized as follow:

1) The reference point of a discourse can be explicitly identified by using temporal conjunctions and frame adverbial phrases.

2) In the case that there is no explicitly mentioned about reference point, the reference point of a discourse can be shifted by:

a. the Aktionsart of a main clause; accomplishments and achievements introduce new reference points, while states, activities and events described in the progressive do not.

b. the use of temporal conjunctions.

c. the use of flexible anchoring adverbials and dependent adverbials.

3) As shifters of reference points, temporal conjunctions, flexible anchoring adverbials and dependent adverbials share the anaphoric function of tense morphemes; they depend for their interpretation on a reference point previously established in the discourse.

In spatiotemporal zoning, a new zone will be created when new event entered to the discourse happened in time frame different from predecessor events. Specifically say, movement of events' reference time can, but not necessary to, cause a newly instantiated zone. The basic strategy for zone instantiation will be described next in section 6.3. Here I will mention about the interpretation of reference time in the discourse.

The temporal discourse interpretation principle (TDIP) introduced by Dowty [25] gives a basic idea of the way to interpret reference time in a discourse. The definition of TDIP is shown below.

Temporal discourse interpretation principle (TDIP)

Definition: Given a sequence of sentences S_1, S_2, …, S_n to be interpreted as a narrative discourse, the reference time of each sentence S_i (for i such that $1 < i \leq n$) is interpreted to be:

a) A time consistent with the definite time adverbials in S_i, if there are any;

b) Otherwise, a time which immediately follows the reference time of the previous sentence S_{i-1}.

The intent of the phrase "immediately follow" in the TDIP is that the reference time of the sentence S_i is to be the very next event or state (or narrator's perception of a state) of significance to the narrative.

2.4.3 Temporal move of reference point

As the narrative progress, temporal reference point (or "temporal focus" [28]) changes its position in time. The movement of reference point or temporal focus from one clause/sentence of the narratives to the next is sometimes called as the narrative move [28].

According to Nakhimovsky [28], there are two kinds of narrative moves, which are called as micro-moves and macro-moves. He illustrated 3 examples, which are shown below, to reflect two kinds of narrative move.

(1) **a.** John entered the president's office. **b.** The president got up.

(2) **a.** Gradually, Harvey began to yield the details of his crime, prodded by the persistent questions of the investigator. **b.** He arrived at the bank at 4 p.m. dressed as a postal worker.

(3) **a.** Harley and Phoebe had been sent by their mother to fix the tail valve of the windmill. **b.** In the great expanse of the prairie where they lived, the high tower of the windmill was the only real landmark.

Example (1) shows the temporal-based micro-move of the narrative. In this example, the events progressed orderly within the same narrative unit.

In (2), there is the shift of temporal reference point from event in (a) to an earlier event in (b), over a considerable time interval. This kind of temporal shift signals the beginning of a new discourse segment, as well as temporal reference point.

In (3), Sentence (b) introduces a drastic change in time granularity, rather than movement of temporal reference point. The time granularity of the entire sequence of events of (a) is within a day or two. The time scale of sentence (b), indicated by the "where they lived" clause and the lifetime of a windmill is years or decades. This time-scale change also signals the beginning of a new discourse segment in sentence (b).

The theories mentioned above in this section were a foundation for temporal attribute design in the spatiotemporal framework; both temporal attribute and zone manipulation. The idea of the moving of temporal reference point is important in spatiotemporal zoning in that, the movement (macro-move) of narrative's temporal reference point can be considered as an instantiation point of new spatiotemporal zone.

2.5 Characteristics of news report on disease outbreaks

In order to effectively design the spatiotemporal zoning scheme, it is necessary to study the characteristics of the disease outbreak news report.

Outbreak news reports are news stories that aim at reporting events that describe the spread or occurrence of infectious diseases. They can report either about single situation, i.e. macro event [91], or the succession of the connected, ongoing situations.

News stories relate more than one event (the definition of event here cf. to [100]) and there must be some relevant connection between these events.

News reports on disease outbreaks are the same as other news stories, where they always have an abstract at the beginning, usually consists of a headline and one or two lead sentences. This part generally briefly states what happened, together with the location and time of the story's occurrence. In the abstract part, information about location of the situation is usually given on a broad scale, such as the national or provincial level.

The structure of outbreak news report is quite complex, intertwining and dispersing descriptive of orientation [101] with story events throughout the narrative. The orientation of news stories provides background information about time, place, as well as person and their activities or the situation. News stories generally contain orienting information, but rarely have a separate orientation section at the beginning; it usually can be found dispersedly throughout the news story. The structure of news stories is complicated mainly due to the requirement of brevity.

The dispersion of the orientating information may be contributed from several reasons.

- News stories may describe more than one outbreak event, place or time. Also, they may relate to more than one person. So it may be necessary to introduce a new orientation parts throughout the story.

- News stories are sometimes in the form of an update report describing the evolution of the outbreak characteristics.

- The requirement of the compactness causes the orientation to be conveyed in subordination in new stories and embedding in the narration of the complicating events. The orientation is generally expressed in relative clauses, and adverbial phrases or clauses of time, place, manner, etc. [101]

Initial news reports or real-time update reports are generally condensed and include only the gist of information regarding to current situation, such as the disease, location, victims, and so forth.

Follow-up reports or daily reports, however, can be long and composed of various views on a situation. They may consist of updated information, as well as previously

reported data about the current situation. They can also include basic knowledge about the disease, control measures for the situation, personal information about victims, or suggestions to residents in handling the spread of the disease.

Another characteristic of news in terms of spatial aspects is worth mentioning. In news reports, when no geographical location or well-known location is reported in the story, the situation typically occurred in the same geographical area as that of the news agency.

Chapter 3

Spatiotemporal Zoning

Based on the requirement of the target applications, shaped with the basic theories previously studied, in this chapter, a novel scheme, namely spatiotemporal zoning, is introduced. This scheme attempts to recognize geo-temporal information of an event at the finest level of granularity available in text. This scheme was proposed as another means to mitigate the limitations of current report-based surveillance systems by allowing for a fine-grained understanding of the spatiotemporal information of events.

The spatiotemporal zoning scheme is represented in the form of a mark-up language that describes the spatial and temporal information of the textual content. Generally, the purpose of mark-up languages is to provide an inter-changeable format for electronic documents, where text content is enclosed by structured text descriptions, called tags. Tags give clear and concise information about the data which they enclose. Within tags, attributes can be given in order to provide additional information about the data. Since the structure of mark-up language must be defined a priori, computer programs can automatically parse marked-up documents and understand the content easily.

The organization of this chapter is as follow. The first section defines the events considered in the framework and gives a concise description of spatiotemporal zoning. Next, the classification of news content according to its spatial and temporal information is described. The issues faced in the spatiotemporal framework design are discussed next. Finally, the spatiotemporal zoning scheme is described in detail.

3.1 Definition of events

Since this thesis is dealing with analysis of the times and places of events reported in natural language text, it is necessary to explicitly specify the definition of events. Here, the definition of an event follows the definition used in the TimeML framework [98], where “event” is a blanket term for situations that happen or occur. Events are considered as predicates describing the states or circumstances in which something

changes, obtains, or holds true, and which might need to be located in time. Linguistically, an event is typically defined as a single clause that contains one predicate (i.e. verb) and its arguments (e.g. subject or object).

In spatiotemporal framework, events may be expressed by the following means:

1) Tensed or un-tensed verbs, such as "die", "occur", and "spread".

2) Verb phrases in the form: 'to be + certain sets of adjectives', such as "(is) underway" and "(was) ill".

3) Verb phrases in the form: 'to be + prepositional phrases', such as "(is) on board", "(is) under construction", and "(is) in progress". Prepositional phrases that indicate locations, such as "in Indonesia", are also considered as events in spatiotemporal framework.

In the rest of this thesis, "event" or "event expression" means a linguistic constituent consisting of a sentence, finite clause, non-finite clause, or phrase that contains a single event, as in the following example:

> Seventeen people **have died** and 41 **have been admitted to hospitals** in Sichuan, China, **suffering** from an undiagnosed disease.

Expressions marked in bold face represent three different events.

3.2 Spatiotemporal Zoning: Task definition

The objective of the spatiotemporal zoning scheme is to enable language technology software to partition text into set of coherent text 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 homogeneous time frame. Here, a homogeneous time frame means that events or actions mentioned in the text segment overlap in time, occurring either continually or sequentially. The text capture shown in figure 3-1 below is an example of spatiotemporal zoning.

```
<ZONE id=1 Type="EVENT_Report" Location_id="2&3&4" Anchor_Val="2006-11-21"
Val="Past_REF" STime="2006-9-1" STime_Dir="AS_OF" ETime="2006-11-8"
ETime_Dir="AS_OF">
From 1 September to 8 November 2006, 16 deaths of meningococcal disease have been
reported in Greater <NAME cl="LOCATION" id=2>Yei</NAME> County, Central <NAME
cl="LOCATION" id=3>Equatorial</NAME> State of South <NAME cl="LOCATION"
id=4>Sudan</NAME>.
</ZONE>

<ZONE id=2 Type="EVENT_Normal" Location_id="2" Anchor_Val="2006-11-21"
Val="Past_REF" STime="2006-W44" STime_Dir="AS_OF" ETime="2006-W44"
ETime_Dir="AS_OF">
The epidemic threshold was crossed in this county during the last week of October.
</ZONE>
```

Figure 3-1: Text capture of spatiotemporal zoning in a news report

Text capture shown in figure 3-1 was marked-up with spatiotemporal zone according to the spatiotemporal zoning specification. The first zone is report zone consists of one event, which is *reported*. This event occurred in Yei County, Central Equatorial, in Sudan from 1 September to 8 November 2006. These spatial and temporal information are represented in the zone's Location_id, STime, and ETime attributes, respectively. The second zone also consists of one event, which is *crossed*. This event is annotated as occurred in Yei County, in the last week of October 2006, according to information available in the news report. The detail of each zone attribute is described later in this chapter.

3.3 Zone attribute: Class of news content

There are some text segments whose content cannot be anchored along a timeline. Since spatiotemporal zoning deals with analysis of events' temporal and spatial information, the capability to distinguish groups of expressions from one another is necessary. In terms of spatiotemporal characteristics, news content can be classified into three main groups.

The first group is text segments that contain "happening events". News content in this class is composed of events that truly occur in the world and hence can be located along a timeline. These events' occurrences could be in the past or present, or will definitely occur in the future. The second group of news content is "hypothetical events". Text segments falling in this group consist of events that may or may not happen. These

events are usually based on an expectation, prediction, belief, or thought. The last group is “information”. This group contains either generic events that cannot be specifically located in time, or eternal truths. Text segments classified into the information group are usually given as basic knowledge, such as general information about a disease, for news readers.

Regarding utility in a health surveillance system, news content in these classes also contributes a greater or lesser degree of usefulness to the situation analysis task. Text segments that contain happening events are usually the main interest of people who want to learn about newly occurring situations or track the continuation of any situation. Hypothetical events are usually not considered the main focus of situation analysts, since they are usually based on a prediction or personal opinion. These events might be useful, however, as information for prevention or a control strategy. Text segments that explain basic knowledge related to outbreaks, such as details about a disease or pathogen, are generally not regarded as significant to the event tracking or event detection task.

Since text segments of each class possess different levels of significance from the situation analysis point of view, I believe that it would be useful to include this information in the spatiotemporal zoning framework, as well. Therefore, this work classified text content into 3 main classes, which are general information, hypothetical event, and Temporally-locatable event. The following subsections describe the details and characteristics of each class.

3.3.1 Generic information

Generic information is usually non-eventive expressions, events that can not be positioned in space or time, general knowledge that is always true, or generic events [98]. The following are examples of generic information:

1) General knowledge that is always true or events that cannot be located in time. For example:

(1) Chikungunya is spread when tiger mosquitos drink blood from an infected person and, if conditions are right, pass the virus on when they bite again.

2) Imperative and interrogative sentences, as well as recommendations, suggestions, and requests. For example:

(2) Residents are recommended to stay away from the facility for one month.

(3) Students with symptoms should stay out of school until they have taken antibiotics for at least five days.

3) Sentences whose subjects are linked to their predicates (e.g., characteristics, attribute, etc.) via a copula verb. For example:

(4) The victim is a 12-year-old boy.

(5) He is a resident of Boyolali district.

Informative-type expressions in the second and third groups usually convey information about the current situation, such as the details of victims, control measures, and so forth. In contrast, expressions in the first group, i.e., general knowledge, only provide basic information to readers.

3.3.2 Hypothetical event

Hypothetical events are those that are alternative or occur in other possible worlds. Events in this group represent only the perspective or anticipation of the speaker. While hypothetical events may or may not happen, forthcoming events are those that, without any unexpected circumstances, will definitely occur in the future, such as events that are planned. Generally, hypothetical events are the following:

1) Events introduced by verbs such as hope, believe, think, expect, and so forth. For example:

(1) Health agencies in Tripura were expected to launch culling operations later on Monday.

2) Events marked with a modal such as “may” or “might”. For example:

(2) Some of the patients might need to remain on the respirators for a couple of months.

3) Events introduced in a conditional sentence, i.e. in an *if*-clause. For example:

(3) If the virus mutates it could create a pandemic that would kill millions of people.

3.3.3 Temporally-locatable event

Temporally-locatable events are those that have happened, are ongoing, or will definitely happen. Among linguistic expressions that represent Temporally-locatable events, there is a certain set of events that are worth mentioning. These events have a communicative function, i.e. represented by verbs of communication [102], and I refer to them as reporting events. From a grammatical perspective, the timing of reporting events has an

influence on the temporal interpretation of events in the scope of a quoted speech. This influence is apparent in direct speech construction, where the time of an event inside a quotation is interpreted in terms of both the tense and time of the reporting verb and the event's own tense. Given this characteristic, I believe that it is advantageous to separate reporting events from other occurrence events [91]. For spatiotemporal framework, the temporally-locatable events are further classified into two subclasses: reporting events, and normal events.

3.3.3.1 Reporting event

Reporting events [91] describe the action of a person or organization declaring something, narrating an event, informing about an event, and so forth [98]. This type of event is used to give information about what people or organizations say or think. Reporting events are usually expressed by reporting verbs, such as "say", "tell", "announce", and "report". Examples of reporting events are shown below:

(1) The ministry **said** the boy might have been infected by sick chickens near his home.

(2) "It's very important to test the vaccine on humans and to produce it," Van **said**.

(3) At least 15 children had died in the outbreak, Health Department Director General O.P. Singh **confirmed**.

3.3.3.2 Normal event

Normal events are Temporally-locatable events that are not reporting events. Statistical investigation showed that more than 50% of the event expressions in the representative corpus are in the normal class. Examples of normal events are the following:

(1) A total of 14 of the 19 districts in the state, including Murshidabad, **had been affected**.

(2) Five days after **returning** to her hometown of Khon Kaen, she **fell ill** with Sars-like symptoms.

Normal events also include fluents, which are used for representing a specific property of an object that has time duration. For example, "All the patients are stable", "The girl was in a serious condition".

3.4 Issues in Spatiotemporal zoning framework design

This section discussed about some issues in the designing of the framework regarding to the zone attributes.

3.4.1 Issue in news content classification

On the surface level, one linguistic expression (i.e. clause or phrase consisting of one verb) could conform to the definition of more than one content class, such as in the following example:

(1) Rabies **attacks** the brain and nervous system.

(2) If **untreated**, meningitis **is** a potentially serious condition owing to the proximity of the inflammation to the brain and spinal cord and **can lead** to death.

In the example (1), the expression belongs to the information class while its surface form conforms to the specification of the normal class. The example (2) represents the case of information-class expression whose surface form is similar to hypothetical-class expression.

However, in this framework, each linguistic expression in text is mutually-exclusively classified into only one content class. The classification is based on a role that the linguistic expression plays in the document. Generally, the role of linguistic expression is indicated by the expression itself together with contextual information around it. For example:

(3) Pertussis **begins with** cold symptoms as sneezing, a runny nose, a low-grade fever, but the cough **becomes** more violent and **may lead** to vomiting.

In the example above, the surface form of the expression “may lead” indicates the hypothetical class. However, when considered together with its surrounding context, the role of the expression is to provide basic knowledge about the disease, and should be classified as information class.

3.4.2 Issue about temporal attribute

The temporal aspects of spatiotemporal zones are somewhat more complicated than the other attributes. Since a zone is defined as a text segment that consists of a group of

events that occur in the same span of time, representing the occurrence time of events in a zone with one attribute may not always be appropriate. One of the most obvious examples is a news report about continuation of events over a certain period, as in the following sentence captured from the news article:

> From 1 September to 8 November 2006, 16 deaths of meningococcal disease have been reported in Greater Yei County, Central Equatorial State of South Sudan.

To enable the scheme to handle these cases, I regard the temporal attribute of the zone as a period with a starting time and an ending time. Hence, instead of defining the event time as one temporal attribute, two temporal attributes are introduced in order to indicate the starting time, STIME, and the ending time, ETIME, of the event occurrence period.

Another issue to consider is the relation between events in the zone and time. As reported previously [103], events and time can exhibit various relations, e.g., before, after, simultaneous, and so forth, as shown in the example below:

> The man was declared brain dead on Aug. 26, three days after suffering a serious head injury.

Therefore, it is necessary for the scheme to provide a means to reflect the temporal relations between events in a zone and the starting and ending times of events' occurrence. In spatiotemporal zoning scheme, I introduce two attributes to express such temporal relations: STIME_Dir to indicate the temporal relation between events in the zone and the starting time of the occurrence period; and ETIME_Dir to indicate the temporal relation between events in the zone and the ending time of the occurrence period.

Another important element is the reference time. Generally, the presence of reference time is not significant when the readers know an event's absolute time. I often find cases, however, in which there not enough information is available to infer the absolute event time. These cases are usually those in which the occurrence time is represented by means of verb tense. With only the tense available as temporal information, only an approximate occurrence time relative to a certain reference point can be determined, as in the following example:

> At least 45 people have died of malaria in Jalpaiguri and Coochbehar Districts of North Bengal, senior health department officials said on Thursday.

In the above sentence, we only know that the event "died" started to occur at some time before the utterance time and continued to occur until then, at least. In these situations, the reference time plays an important role in temporal interpretation. Hence, it is necessary to include the reference time as a temporal attribute of a spatiotemporal zone.

Application of temporal attributes

While spatial information is an attribute that is inherent to any type of zone, temporal information cannot be considered in that way. Temporal attributes can be applied to the text contents that can be located in time only, i.e. either normal or reporting events.

3.4.3 Temporal granularity in health news events

In outbreak news reports, temporal expressions at the level of time-of-day granularity are rarely reported. On the contrary, we usually find that temporal expressions in outbreak reports are at the level of a 'day' or a coarser period, such as a week, month, or year. In terms of requirements, organization of news reports in health surveillance systems with regard to time is done at the level of a day, i.e., news is grouped and presented on a daily basis. Given these considerations, in the spatiotemporal zoning framework, the temporal attributes are specified with 'day' granularity.

3.4.4 Spatial granularity in health news events

The spatial attribute of the event can be selected from any expression considered as location entity according to the BioCaster named entity annotation specification [104]. In BioCaster project, location entity is the expression that absolutely refers to the politically or geographically defined location at any granularity. In spatiotemporal zoning, locations to be selected as the places where the events occurred are preferred to be those with the finest level of granularity according to the information available in text.

3.5 Spatiotemporal zoning annotation scheme

This section introduces the details of the spatiotemporal zoning scheme. Nine attributes for zone annotation are summarized in table 3-1.

Table 3-1: Zone attribute for spatiotemporal annotation

Attribute		Description	Value
ID		Represents zone's ID	Number
EVENT TYPE		Indicates the type of contents in a zone	"Information", "Event_Hypothetical", "Event_Report", "Event_Normal"
LOCATION		Specifies the geographical location where the events in the zone happened	The location attribute can be any politically or geographically defined location. To be specific, any location expression that is considered annotatable according to the specification used in the BioCaster system [104] can be a location attribute.
Temporal attribute	ANCHOR_VAL	Indicates a reference time used for interpretation of other temporal attributes	The default value is the document date or news report date. If the events in the zone are in the scope of direct speech, ANCHOR_VAL is the date of the reporting event.
	VAL	Indicates a relative time with regard to the value of ANCHOR_VAL, at which, according to the available textual information, the events in the zone hold true or happened	PRESENT_REF for a present event, PAST_REF for a past event, and FUTURE_REF for a future event
	STIME	Indicates the (approximate) starting time of the events in the zone	The normalized form of a temporal expression reported in the text, or the value of "PAST", "PRESENT", or "FUTURE"
	ETIME	Indicates the (approximate) ending time of the events in the zone	
	STIME_DIR	Indicates relative direction or orientation between the value of STIME and the events in the zone	"AS_OF", "BEFORE", "AFTER"
	ETIME_DIR	Indicates relative direction or orientation between the value of ETIME and the events in the zone	

3.5.1 Zone attributes

The detail of each zone attribute is explained in detail in this subsection. There are 4 main attributes, which are zone id, zone type, spatial attribute, and temporal attribute. The temporal attribute is further separated in to 6 attributes, which are reference time point, relative time with regard to reference time, absolute starting/ending time, relative direction between events in the zone and starting/ending time. In the end of this subsection, the methodology for annotating temporal attribute is exemplified.

ID: Zone identification attribute

This attribute represents a zone's ID.

TYPE: Zone type attribute

This attribute indicates the type of contents in a zone. There are four values for the TYPE attribute. These values are defined according to the classes of text content. They are: "Event_Info" for the information type, "Event_Hypothetical" for the hypothetical type, "Event_Report" for the Temporally-locatable reporting type, and "Event_Normal" for the Temporally-locatable normal type.

As mentioned earlier, information or hypothetical events cannot be located along a timeline. As a result, text contents marked with the Event_Info or Event_Hypothetical value will have no temporal attributes marked in the zone.

LOCATION: A location attribute

The location attribute specifies the geographical location where the events in a zone happened. The location attribute can be a geographical location at any granularity available in a text.

ANCHOR_VAL: Reference time attribute

The ANCHOR_VAL attribute is introduced with the purpose of giving a reference time, which is used for interpretation of other temporal attributes. The ANCHOR_VAL attribute consists of a normalized form of an anchoring date.

Generally, the default value of ANCHOR_VAL is the document date or news report date. In the case of direct speech construction, the timing of events in quoted speech is interpreted with regard to the time of speaking. Hence, if events to be annotated are in the scope of direct speech for a reporting event, the date of the reporting event is selected as the value of ANCHOR_VAL.

VAL: Relative time attribute

The value of the VAL attribute indicates a relative time with regard to the value in ANCHOR_VAL at which, based on the available textual information, the event in focus hold true or happened. For example, if the events in the zone occurred in the past, then the VAL attribute is "past", but if the events started occurring in the past and have continued until the present time, the VAL attribute is considered "present".

Hence, there are three possible values for the VAL attribute: PRESENT_REF for present events, PAST_REF for past events, and FUTURE_REF for future events.

STIME: Starting time attribute

STIME indicates the (approximate) starting time of the annotated events. The value in STIME is the normalized form of a temporal expression based on the information available in the text. If there is no explicit information indicating the starting time of events in the zone, however, the value in STIME can be PAST, PRESENT, or FUTURE, in relation to the value of ANCHOR_VAL.

ETIME: Ending time attribute

ETIME is the same as STIME, with the difference that ETIME indicates the approximate ending time of the annotated events.

STIME_DIR: Relative direction with regards to starting time

In many cases, an event time is reported by using a preposition to indicate a temporal relation between the time that the event happened and a temporal expression or another event. In this circumstance, neglecting the existing of the preposition would result in the loss of detailed information for locating events along a timeline. In an attempt to reserve all explicit information as much as possible, it is necessary introduce temporal attributes

that reflect this type of temporal relation, namely STIME_DIR and ETIME_DIR, which are explained next.

The STIME_DIR attribute represents the relative direction or orientation between the value of STIME and the events in the zone. In the TimeML framework, there are 13 temporal relations between events and temporal expressions or other events. These relations, however, are very detailed and quite complex. To eliminate unnecessary complexity, I decided to group these relations together and classified them into three main classes, which correspond to the possible values of STIME_DIR.

The value of STIME_DIR can be any of the following:

1) AS_OF

This class consists of the following types of temporal relations defined in TimeML: “simultaneous”, “including”, “being included”, “during”, ”being held during”, “beginning”, “begun by”, “ending”, “end by”. The AS_OF relation is comparable to the OVERLAP relation in the SemEval-2007 TempEval task [33].

2) BEFORE

This class consists of the following types of temporal relations defined in TimeML: “before”, and “immediately before”.

3) AFTER

This class consists of the following types of temporal relations defined in TimeML: “after”, and “immediately after”.

ETIME_DIR: Relative direction with regard to ending time

ETIME_DIR is the same as STIME_DIR, except that it represents the temporal relationship between the value of ETIME and the events in the zone.

Temporal attribute annotation example

The example below illustrates the selection of each temporal attribute in zone annotation, for a publication date of April 14, 2005.

> Two of the cases were recently detected, between 2 and 8 April, in Hung Yen and Ha Tay Provinces, respectively.

In the above example, the VAL attribute is selected as PAST, since the event "were detected" was completed by April 8, before the publication date. The temporal expression in this sentence indicates that the event occurred at some point in time from 2 to 8 April. According to this, STIME and ETIME are annotated as "April 2, 2005" and "April 8, 2005" respectively, with both STIME_Dir and ETIME_Dir selected as "AS_OF".

Relation between STIME and ETIME

In many cases, the value of STIME attribute, t_s, and ETIME attribute, t_e, can be the same. However, t_s, is always considered to appear before or the same time as t_e; $t_s \leq t_e$. For example, t_s and t_e can be the same day, but different time (e.g. 12.00 and 13.00), or the same month but can be the different days, etc.

If the zone, z_i, consists of more than 1 non-parallel events, then $t_s < t_e$.

If the zone, z_i, consists of single punctual event or parallel punctual events, then $t_s = t_e$.

BNF for Zone tag

Attributes ::= ID TYPE TEMPORAL_INFO LOCATION

TEMPORAL_INFO ::= ANCHOR_VAL VAL STIME STIME_DIR ETIME ETIME_DIR

ID ::= id

TYPE ::= 'Event_Hypothetical' | 'Event_Info' | 'Event_Normal' | 'Event_Report'

ANCHOR_VAL ::= CDATA

LOCATION ::= LocEntity | LocSeq

ANCHOR_VAL ::= document_creation_time | news_report_time | utterance_time

VAL ::= 'PRESENT_REF' | 'PAST_REF' | 'FUTURE_REF'

STIME ::= ISO_Time | 'PAST' | 'PRESENT' | 'FUTURE'

ETIME ::= ISO_Time | 'PAST' | 'PRESENT' | 'FUTURE'

STIME_DIR ::= 'AS_OF' | 'BEFORE' | 'AFTER'

ETIME_DIR ::= 'AS_OF' | 'BEFORE' | 'AFTER'

LocEntity ::= location NE | compound location NE

LocSeq ::= Loc_Entity&Loc_Entity(&Loc_Entity)+

3.5.2 Metadata

The metadata section is introduced into the spatiotemporal zoning framework with the purpose of providing extra information related to the zone annotation task, including both the temporal and the spatial aspect. Currently, the metadata section contains the information as described below.

3.5.2.1 Temporal-related Metadata

The temporal metadata section of an annotated document provides information related to temporal attributes. The temporal metadata consists of the following:

1) News publication date

The news publication date is introduced in the metadata section since it is regarded as the default value for the anchor date in the zone annotation task

2) The ISO normalized form of each temporal entity marked in the text

This second part of temporal metadata consists of the ISO normalized form of each temporal entity marked in the text. The purpose of this metadata is to canonicalize relative temporal expressions in news articles, such as “yesterday”, “today,” and “tomorrow,” to absolute times, as well as to convert each temporal expression to the same format. Since the smallest unit of time used for zone annotation is a ‘day’, temporal entities with finer granularity than the ‘day’ level will be associated with the normalized forms of their dates.

The text captured in figure 3-1 is an example of temporal metadata. The value marked up by Anchor_Date is the news publication date. In the TIME_Norm field, the ‘time_id’ is the ID of each temporal expression appearing in the text, and ‘norm’ is the ISO normalized form of that temporal expression.

```
<TEMPORAL_INFO>
<Anchor_Date>2004-1-31</Anchor_Date>
<TIME_Norm time_id="1" norm="2004-1-31">
<TIME_Norm time_id="2" norm="2003-12-16">
</TEMPORAL_INFO>
```

Figure 3-1: Temporal metadata in the spatiotemporal zoning scheme

3.5.2.2 Location-related Metadata

The location metadata consists of two parts. The first part provides information about the site of a news agency. The second part provides relations between each location appearing in text.

1) Location of news agency

Generally, a news agency's location is specified at the country level. If the news agency is local to a lower level of administration, however, a geographical location at the city or province level is used. This information can be used as the default value of an event's location, since observation indicates that many news agencies usually omit the locations of reported events when they are local to the agencies' locations.

2) Geographical relation

This part of location metadata provides information about the relations holding between location expressions that are often found in text. Currently, two relations are considered in this scheme: "IS_A", and "PART_OF".

"IS_A(A,B)" indicates that locations A and B are geographically the same. For example, IS_A("USA", "United States of America") indicates that the two location names both refer to the same geographical area.

"PART_OF(A,B)" indicates that location A is located in location B. For example, PART_OF(Tokyo, Japan) indicates that Tokyo is located in Japan.

3.5.3 Unit of annotation

As mentioned earlier, events can be expressed by various kinds of linguistic component. It is likely that a single sentence may consist of multiple constituents expressing spatiotemporal-separated events. The data observation confirmed my hypothesis. From data observation, there are many cases where a single sentence consists of constituents (e.g. clauses or phrases) pertaining to different event type, or mentioning about events that occurred in different place or time. For example:

> The father became ill on 2 July with fever, mild cold, then coughed and was taken to the district hospital on 7 July where he died 10 days after onset.

In such a case, giving a single annotation to the whole sentence would cause inappropriate aggregation of the spatiotemporal attributes of independent or irrelevant events into a single zone. As a consequence, detailed information, which is possibly important, may be made obscured. The clause-level annotation is considered to be more complicated to the annotator than the sentence-level annotation. However, as a trade-off for the purpose of retaining the detailed information as much as possible, it should be more appropriate to define the annotation in clause-level. Grammatically, a clause is an expression consisting of a subject and a predicate. It is a linguistically stable and easily recognizable unit of annotation for the general readers. Therefore the framework proposed in this thesis decided on the clause-level annotation.

After investigating a number of news reports, however, I found that events conveyed in different types of clauses contribute different levels of importance to the main situation reported in news. For example, noun-modifying clauses (e.g., relative clauses, noun-modifying non-finite clauses, etc.) usually give supplemental information, including possibly unimportant past events. Moreover, these clauses can be considered as the events that are not in focus from the reporter's perspective. Therefore, it necessary to specify syntactic terms the constituent types that qualify for an independent zone. With this restriction, spatiotemporal annotation can avoid over-generation of small, scattered zones, which would lead to results that are too complex and unnecessarily detailed. Table 3-2 lists the allowed constituent types. In the table, the square brackets in the examples indicate zone boundaries. Here, A and B are considered as clauses representing events that are either spatiotemporally different or different in type.

For the spatiotemporal zoning framework, in general, an annotator processes text sentence by sentence. If all events conveyed in a sentence are temporally coherent events that occurred at the same geographical location, the whole sentence is annotated with a single zone, with attribute values derived from the events. If the events expressed in a sentence appear to occur at different times and/or locations, however, a new zone is instantiated. If an adjacent sentence (or clause) contains spatiotemporally similar events, then it is annotated together in the same zone as that of the first sentence or clause.

Table 3-2: Constituent types qualifying for an independent zone annotation

1) A sequence of sentences
2) A sentence
3) Coordinating clauses For example: [A] [and B]
4) Subordinate clause [105] - Subordinate clause introduced with a subordination marker (e.g., that, whether, and if) For example: [A said that [B].] - Subordinate clause introduced with a word functioning as the head of the constituent. For example: [A] [when B.]
5) Non-finite clause [105] - Infinitive For example: [A [to B].] Note: If an event expressed by an infinitive cannot be located in time, such as an event expressing a purpose, goal, or intention, I do not consider it to qualify for being in an independent zone. - Gerund For example: [A, [saying …].]

3.5.4 Nested annotation

Generally, zones are annotated sequentially. It has been observed, however, that some sentence structures express syntactical dependency between events for which it may be beneficial to allow them to be annotated in a nested manner. The typical cases of nested annotation in a sentence are a subordinate clause or non-finite clause (see Table 3-2), as in the following example:

[He said that [the Indonesia case count has climbed to 226].]

From the data investigation, I found that this nested construction mainly occurs when subordinate or non-finite clauses are the complements of verbs. The most frequent cases are a subordinate clause that is an internal complement of a verb such as “hope” or “expect”, or of a reporting verb such as “said” or “report”. Nested annotation is necessary because of complex sentence structures and also because it provides the possibility of greater flexibility in event analysis. For example, a health surveillance system might pay more attention to the main events in a main zone than to the

complement events in a nested zone, or it might considered the importance or recentness of the main events to proportionally transfer to nested events.

3.5.5 New zone instantiation

In the zone annotation task, the boundary of a spatiotemporal zone is sequentially extended as long as it consists of only a set of expressions (i.e. clauses or sentences) conveying contents in the same class, and/or expressing events occurring at the same location in the same period of time. Whenever two consecutive expressions are considered to be in a different class, or expressing events that occur at different locations or in a new, non-homogeneous time frame, a new zone is initialized with attributes according to its member events. Generally, a set of events is considered to occur in the same homogeneous time frame when each event in the set connects to every other event, either directly or indirectly, through the chain consisting of the 13 relations defined in TimeML, without having a narrative macro-move [28] in the chain. Applying the results of a study [22] on the relation between the usage of temporal markers and topic continuity/discontinuity, I hypothesize that without explicitly mentioning the dates and times of events, differences in event timing are not significant from a reporter's perspective. Without temporal expressions or temporal adverbials, reporters seem to emphasize the continuity of sequentially reported events rather than the timing of each event. Thus, when there is no strong evidence indicating discontinuity of events, I consider these events to occur in the same homogeneous time frame. A separate zone is also created when two textually consecutive events are mentioned to occur with different granularity of time or geographical location. In the case when a group of textually consecutive events occur in the same time frame but at many locations, the events can be annotated within one zone, where all events' locations are identified by the LOCATION attribute.

In the case when two textually consecutive events are reported to occur on consecutive days, in order to keep all detailed information available in a news report, the two events will be annotated in different zones, where the attribute of each zone is encoded from the spatial and temporal of the event which it contains.

Chapter 4

Evaluation of the Spatiotemporal Scheme

In the development of automatic natural language processing systems that involve empirical analysis, annotated corpora have proven themselves to be very important. However, the task of creating large corpora, which generally involves more than one human-annotator, raises concern at least in two respects, which are how to evaluate the annotation scheme and how to assess the reliability of the annotated data. One solution, which has been performed in various computation linguistics tasks, including word sense tagging [106-109], discourse segmentation [94, 95, 110-113], anaphora tagging [114, 115] and text summarization [96, 116], is to show the inter-annotator agreement. In terms of evaluating the validity of the annotation scheme, the resulting reliability indicates how well the annotation scheme captures the truth of the phenomenon being studied [117]. In terms of assessing data quality, data are considered to be reliable if the annotators can be shown to agree, at a certain level, on the annotation task. The agreement on the annotation results allows us to infer that they share the same understanding, and, consequently, I can expect them to perform consistently under this understanding. The reliability of manually annotated data becomes very important especially when they are used to train a system. If the agreement for the annotation is low, then it is likely that the system may replicate the inconsistent behavior of human annotators. As the first step of the development of automatic zone annotation, this chapter discusses the evaluation strategy and numerical evaluation measures to use for the spatiotemporal zoning. Several metrics are used for measuring the agreement. Higher agreement indicates the more reliable of the annotated data and the scheme.

To evaluate the spatiotemporal zoning scheme, an annotation experiment was conducted on 100 pre-selected news articles reporting on disease outbreaks. We were interested in one property, namely, the reproducibility of the annotation scheme. Reproducibility is the extent to which different annotators produce the same annotations. It measures the consistency of shared understanding held by more than one annotator

[118]. Although I did not perform an experiment to evaluate the stability property, i.e., the extent to which one annotator will produce the same annotations at different times, it is commonly assumed that proof of reproducibility implies the scheme's stability [118].

This chapter is organized as follows. First, the data set used in the experiment is explained. Second, the annotation setting, annotators, annotation tool, and the process of annotation, are described. Third, the measures used in the experiment are discussed. Finally, the results of the annotation experiment are qualitatively and quantitatively analyzed.

Note that, although text content in information class is not considered to be an event by definition, to facilitate reference to linguistic expressions to be annotated, in the rest of the thesis, the term "event" is used to cover both actual event expressions and expressions in information class.

4.1 Data set

The corpus for the zone annotation experiment task consisted of 100 news reports about disease outbreak events, randomly selected from the BioCaster gold standard corpus [119]. To mitigate any inconsistency that could occur in the choice of clause boundary, all clauses in the documents in the experimented corpus were assumed to be marked-up. In practice, automatic clausal annotation is a reliable technology [120, 121] with accuracy over 90% average precision/recall.

The first 50 files, denoted as Set1, were annotated by annotators A and B. There were 1086 events to annotate in Set1. The other 50 files, denoted as Set2, were annotated by annotators A and C There were 908 events to annotate in Set2. The number of events to be annotated and the number of sentences in each document set are shown in Table 4-1.

Table 4-1: Data statistics

Corpus	#sentences/clauses/phrases	#events
Set1	808	1086
Set2	518	908

Figures 4-1 and 4-2 show the distributions of documents that contained various numbers of sentences and events, respectively. Here, events were regarded as linguistic expressions that conform to the definition of an event given earlier.

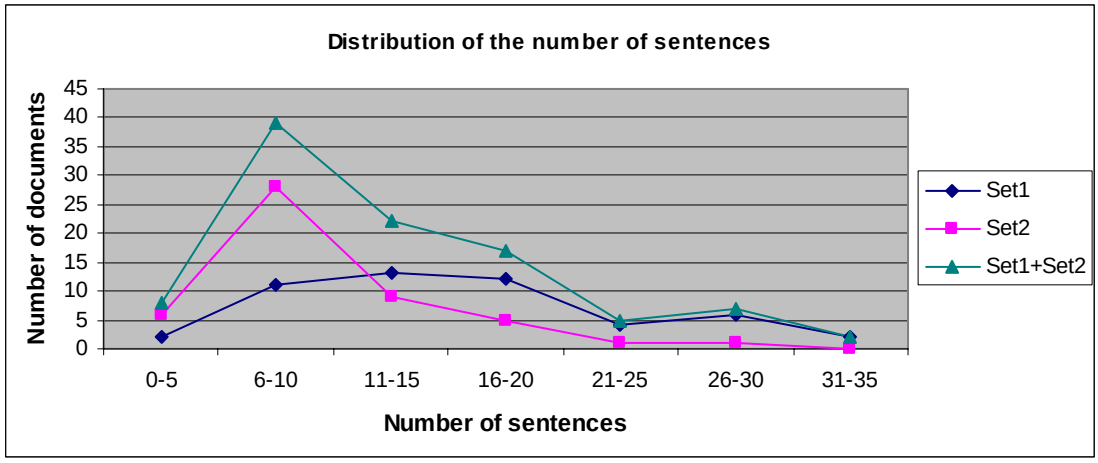

Figure 4-1: Distribution of the number of sentences, including partial sentences

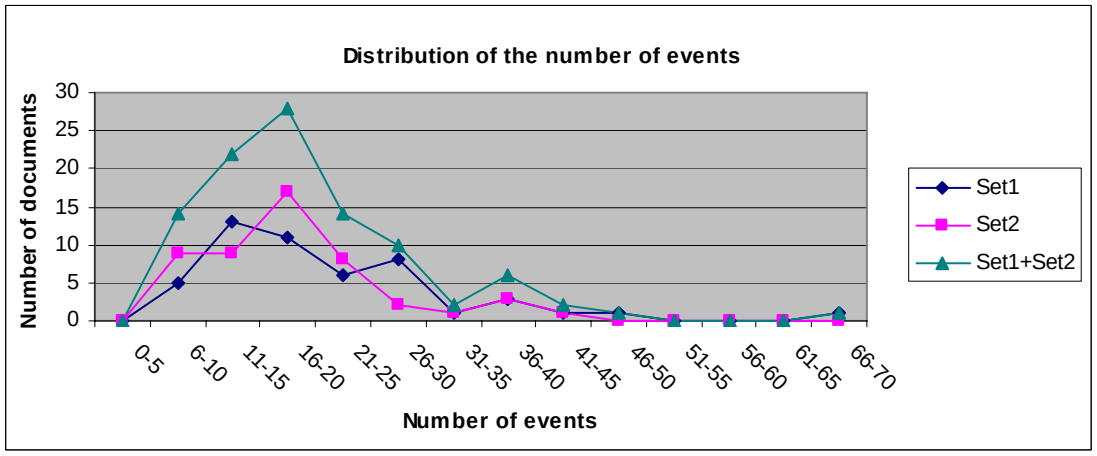

Figure 4-2: Distribution of the number of events, including non-eventive expressions to be annotated

The data represented in figure 4-1 indicates that, overall, documents with a length of 6-10 sentences had the highest proportion. According to figure 5, the majority of the documents in overall consist of 6 to 25 expressions to be annotated.

Figures 4-3 and 4-4 represent the details of outbreak news reports in the experiment corpus in terms of the publication date and affected country respectively. In figure 4-4, the map illustration was created by using Google Maps API[2] for the visualized purpose of location distribution. The chart on the top-left corner of the figure shows the number of documents that report the situation in each country. The corpus covered news articles published from 1996 to 2007. They reported outbreak situations of 44 diseases, on 45

[2] http://maps.google.com

countries world wide. In some articles, one disease outbreak was reported on multiple countries. On the other hand, some articles reported the spreading of multiple diseases within one country.

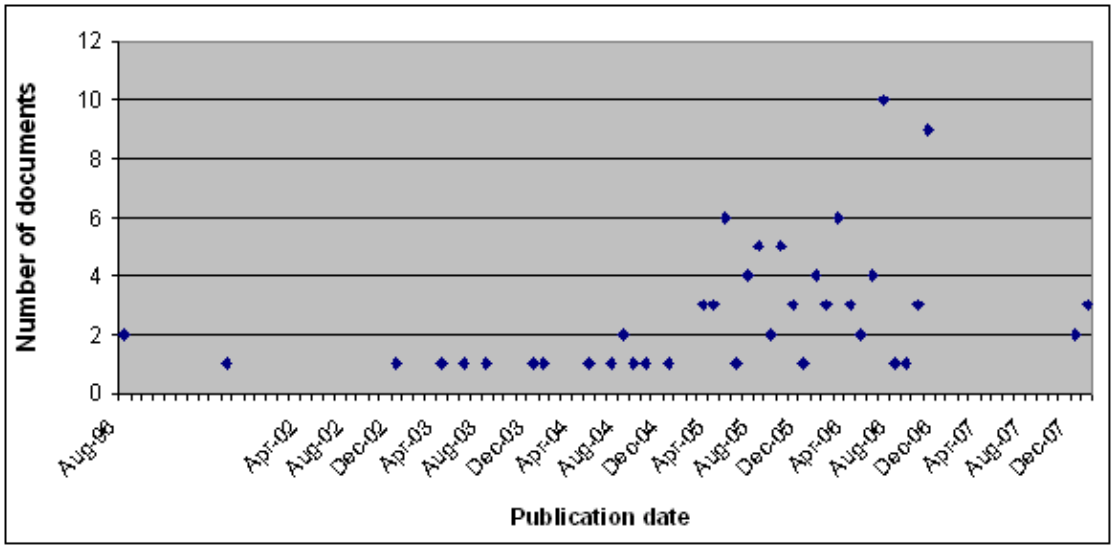

Figure 4-3: Distribution of news articles in the corpus in terms of the publication date

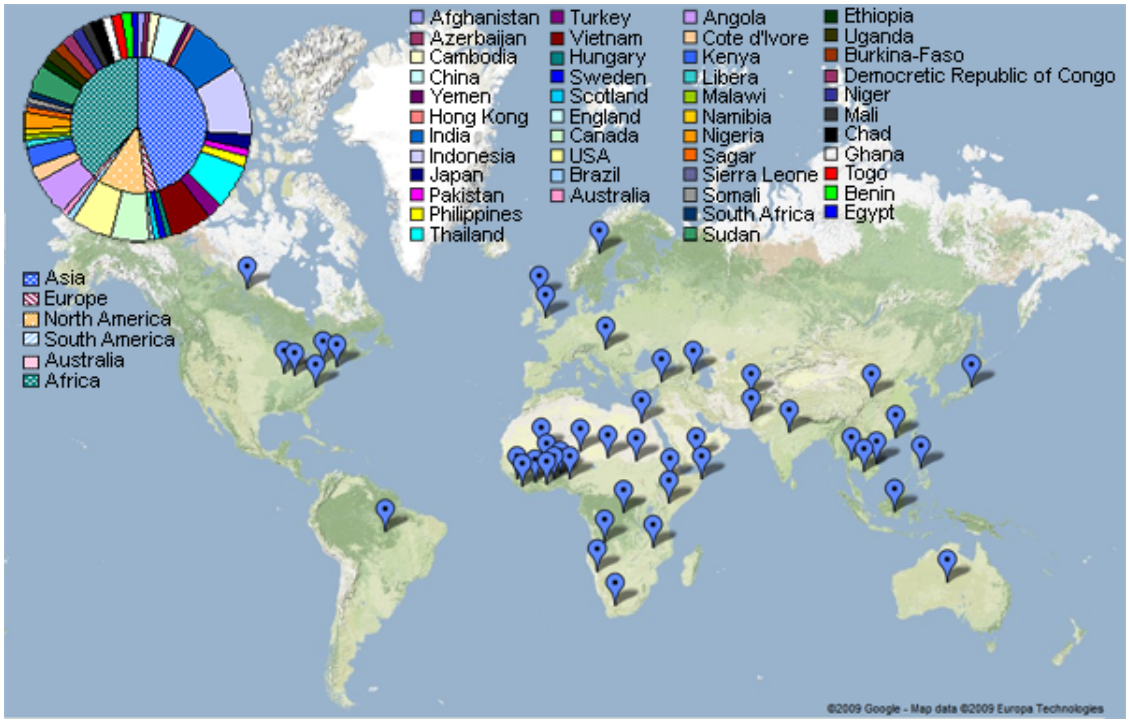

Figure 4-4: Distribution of outbreak situations reported in the corpus, classified in terms of outbreak-affected country

4.2 Annotation setting

In this section, the detail of annotation setting for evaluating the spatiotemporal scheme is described.

4.2.1 Annotators

In the experiment for evaluating the spatiotemporal scheme, there are 3 annotators participated in the annotation task. The first annotator, annotator A, was the author of this thesis. The second annotator, annotator B, holds a Bachelor of Arts degree. The last annotator, annotator C, was a linguist.

4.2.2 Annotation tool

For the annotation task, I developed an annotation tool for spatiotemporal zoning. Since the documents to be annotated could be marked up with various kinds of linguistic information (e.g., named entities), which might cause a visualization problem for an annotator in the zone annotation process, this tool hides these abundant tags from the annotator. It also provides a friendly interface for selecting the boundaries and attributes of each zone. The interface of the annotation tool is shown in figure 4-5.

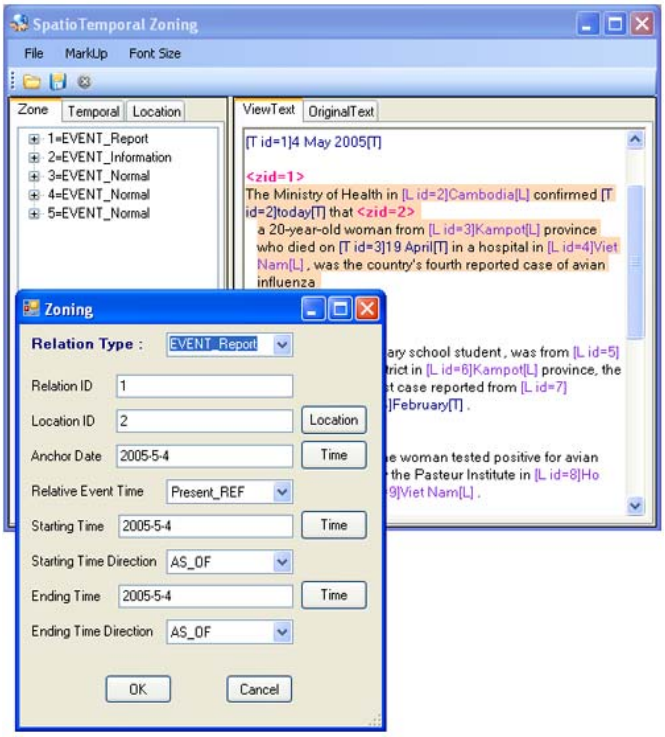

Figure 4-5: Interface of the spatiotemporal zone annotation tool

4.2.3 Annotation guideline design process

In the annotation, it was necessary to have a guideline so the annotators, in some certain levels, have the same understanding and conformably annotate the documents. The annotation task and the design and revision of the annotation scheme were done as a loop process. That is, first, a draft of the annotation scheme was developed in terms of the usefulness of the event-based processing system. Then, news articles were annotated document by document according to this scheme. When there is an issue that was not covered by the original guidelines, further elaborations of the annotation specification were made, and so forth in a continuously looping manner.

4.2.4 Process of annotation experimentation

The annotation experiment consists of 2 phases, which are training phase and annotation phase. In the training phase, the zone annotation guideline was given to the annotators, together with a certain number of examples. After reading the guideline, the annotators were asked to annotate 10 articles selected as training materials, one article at a time. Each file's annotation was followed by a discussion session, with the purposes of clearing up misunderstanding of the guideline, settling disagreements between the annotators, and refining the guideline. After the training, the 50 articles of Set1 were given to annotators A and B, and the 50 articles of Set2 were given to annotators A and C, for annotation.

In the annotation task, the annotators performed full-coverage annotation, meaning that they judged all zone attributes for each span of text containing any single verb. The procedure for annotation was as follows. I asked the annotators to annotate the documents in the same order and send back the marked-up results after every 10 articles. The 10 resulting pairs of articles were thoroughly checked. When apparent mistakes as a result of human error were found, incorrectly annotated articles were sent back to the annotator for re-annotation, without telling the annotator where the mistake was or what the correct annotation should have been. When there was a sign of misunderstanding, or when disagreements occurred repeatedly in the same context, discussion sessions were held in order to settle the confusion. These discussions aimed to reduce the number of undesirable mistakes in future annotation. The numbers of agreements of previously submitted articles, however, were counted before the discussion sessions.

4.3 Agreement measurements

To evaluate the reproducibility of the proposed scheme, the annotation results were analyzed in terms of an inter-annotator agreement measure. For agreement analysis of zone annotation, I considered agreements between annotators in annotating events for each type of zone attribute, which are the zone type, the location attribute, and the temporal attributes (i.e., the starting time, STIME; the ending time, ETIME; and the relative event time, VAL). Since the annotation of each zone attribute can be different in nature, quantitative analysis of annotation agreement on different attributes by only one metric may not be appropriate. In the zone annotation task, the zone type annotation can be viewed as mutually exclusive category assignment, in which a certain value and number of categories are predefined. On the other hand, in annotation of the location and temporal attributes, the annotators can freely select attribute values according to the information available in news text. Therefore, in the quantitative agreement analysis, I used two different statistical measures suiting the different annotation characteristics: the kappa coefficient [122] for zone type annotation, and the percentage agreement for location and temporal attribute annotation.

4.3.1 Kappa

There have been different ways to evaluate agreement between humans for a task characterized as mutually exclusive category assignment. The evaluation matrices use such methods as the percentage agreement and Cohen's kappa coefficient [122]. The kappa coefficient, ***K***, is a statistical measure of inter-annotator agreement for categorical items. It is generally thought to be a better measure of agreement than a simple percentage agreement calculation, since ***K*** takes into account agreement occurring by chance. The equation for ***K*** is the following:

$$\mathrm{K} = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)},$$

where $\Pr(a)$ is the observed agreement among annotators, and $\Pr(e)$ is the hypothetical probability of chance agreement. Regardless to the number of annotators, the number of items to be classified, or the distribution of the categories, $\boldsymbol{K} \leq 0$ means that there is no agreement other than what would be expected by chance, whereas $\boldsymbol{K} = 1$ means that the annotators are in complete agreement.

Since the zone-type annotation could be regarded as mutually exclusive category assignment, kappa was used as an agreement measure for the annotation task.

4.3.2 Percentage agreement

In annotating the location and temporal attributes of marked-up events, the annotators could freely select an event's location as any location name appearing in the news report. Since the nature of the task was not exactly a mutually exclusive classification, the kappa coefficient may not have been totally suitable as an agreement measure. Hence, the simple agreement percentage was used as a measure to show the agreement characteristics between annotators in assigning the location and temporal attributes. The percentage agreement was calculated by the equation below.

$$\mathrm{PA} = \frac{\textit{Number of events in classA, with the same attribute value marked} - \textit{up}}{\textit{Number of events in classA}}$$

4.3.3 Krippendorff's alpha

Krippendorff's alpha (α) [123] is a reliability coefficient for measuring the agreement between observers, coders, judges, or measuring instruments.

α's general form is:

$$\alpha = 1 - \frac{D_o}{D_e}$$

where D_o is the observed disagreement:

$$D_o = \frac{1}{n} \sum_c \sum_k o_{ck\ metric} \delta^2_{ck}$$

and D_e is the disagreement one would expect by chance:

$$D_e = \frac{1}{n(n-1)} \sum_c \sum_k n_c \cdot n_{k\ metric} \delta^2_{ck}$$

When observers perfectly agree, observed disagreement will be $D_o=0$ and $\alpha=1$, which indicates perfect reliability. When observers agree absolutely by chance, the results will be, $D_o=D_e$ and $\alpha=0$, which indicates the absence of reliability.

α can be used to judge a variety of data with the same reliability standards. It can be applied to any number of observers, any number of categories, any metric or level of

measurement, incomplete or missing data, or any sample sizes, not requiring a minimum. In this thesis, Krippendorff's alpha was used for evaluating agreement on event's spatial attribute annotation

4.4 Reliability studies

In this section, the annotation results are analyzed both qualitatively and quantitatively. The quantitative analysis was done for every zone attribute annotation, with the purpose of revealing the reproducibility property of the annotation scheme. The qualitative analysis was done to provide insight into the data in terms of the detailed characteristics of each zone type. This information will be useful for future development of an automatic annotation system.

Since zones are defined to cover spans of text, i.e. multiple consecutive clauses, containing events that occur in the same place at the same time, agreement in zone annotation can occur only when the annotators consistently recognized class, spatial and temporal information for each event. Thus, this study measured inter-annotator agreement on event-based annotation for each zone attribute in order to analyze the details of annotation difficulties.

4.4.1 Qualitative analysis of each zone class

Before reporting the qualitative analysis of the experimental annotation results, it is worth mentioning about the linguistic clues signaling each news content class, i.e., zone class. The following list represents the main linguistic feature types that were found helpful in judging the class of a spatiotemporal zone.

- Lexical item: words and phrases
- verb, verb phrase: including infinitive and gerund
- modal auxiliary (e.g., would, might, can)
- subject of the verb

The subsequent sections present a detailed discussion and analysis of the characteristics of each zone class with respect to these features.

4.4.1.1 Qualitative analysis of Information zone

According to the analysis of the information class, the clauses in this class can express three different characteristics. In order to effectively discuss the qualitative analysis results, I further separate the information class into three groups.

- **Clauses that represent attribute or state of entities**

Clauses in this group represent the attributes or states of entities, such as a person, thing, or organization. These entities appear non-uniformly in news content. The most obvious feature signaling information in this group is the usage of a copula verb to assign a verb's subject to its complement, through verbs such as "be", "seem", and "appear". The following are examples of clauses that fit into this group of the information class:

(1) The fifth victim **was** a 21-year-old woman.

(2) The mother **appeared** to be healthy.

The existence of the auxiliary verb "can" in the sense of indicating capability also signals this group.

(3) It **can cause** serious problems, such as pneumonia and blood poisoning, and can even prove fatal.

- **Clauses that explain about general knowledge or things that are always true**

This group of the information class is the most difficult category to detect, since there are almost no prominent linguistic features to distinguish it from the normal category, which is statistically a major class in news reports. One of the clues that I observed is that the subjects of clauses falling under this group usually refer to classes or conceptual-level entities, instead of to individuals, such as certain types of diseases, viruses, or patients. Examples of clauses regarded as members of this group are the following:

(1) Encephalitis **causes** an inflammation of the brain, **resulting** in brain damage or death.

(2) HTLV-1 **is transmitted** through unprotected sex.

Regarding the location within text, clauses in this group usually appear at or almost at the end of news content. They can occur as part of a person's quoted speech or as a sequence of stand-alone sentences. Moreover, they are generally separated from other types of events. When these clauses occur in a quotation, normally there are only clauses

of this type. When these clauses are stand-alone sentences, they are often positioned in a separate paragraph.

- **Clauses that represent events that can not be spatiotemporally located**

The most apparent clue to signal the existence of this group is the structure of a clause or sentence, which is in the form of an interrogative or an imperative. Clauses that express a suggestive meaning, recommendation, or request are also examples of this group. This type of clause is signaled by the presence of a modal, such as "should" or "must". Examples of sentences in this group include the following:

(1) Was it a migratory bird that got way off track?

(1) Residents **should remember** their basic common-sense health practices.

(2) Provinces with the disease **must immediately set up** steering boards to control the disease.

4.4.1.2 Qualitative analysis of Hypothetical zone

The strong signal of the hypothetical class is the presence of certain words that express the sense of possibility or an expectation. These words are modal auxiliaries, e.g., "would" and "should", or words such as "probably" and "possible", as in the following examples:

(1) A vaccination campaign launched last week **should be able to bring** the outbreak under control.

(2) The virus it had detected at a chicken farm **was probably** not the H5N1 strain dangerous to humans.

The occurrence of certain group of verbs, such as "hope", "expect", and "predict", also indicates the hypothetical class, where the subordinate clauses of these verbs are considered members of the hypothetical class.

(3) In fact we expect that there **will be** more cases.

The modal "can" used in the sense of possibility is another indicator of the hypothetical class.

(4) I am always afraid that any illness **can spread**.

4.4.1.4 Qualitative analysis of Report zone

The strongest clue to signal the reporting class is the verb itself. There is a set of verbs that has a communicative sense, such as "say", "report", and "tell". The existence of these verbs suggests the reporting class.

(1) He **said** that hospitals in this southern Pakistan port city had been put on alert.

(2) She **added** that the severity of the imminent crisis would be worse than previous cases.

Another obvious indicator is when the verb is used as a main verb in direct speech construction, which is marked by the presence of quotation marks, as in the following example:

(3) "There have been no yellow fever reports since 1942 and it has been eradicated since then," he **said**.

Verbs used in this manner are normally regarded as indicating the reporting class.

4.4.1.5 Qualitative analysis of Normal zone

According to the data investigation, the normal class is the predominant class in news reports. It could be considered as the default class for clauses when they do not conform to the characteristics of any of the other classes.

There are no apparent features signaling the normal class. A wide range of verbs are used to express events in this class, as in the following examples:

(1) A 16-year-old Indonesian girl has **died** of bird flu.

(2) The government **transferred** drugs to those regions in order to control the epidemic.

The tense used with verbs in normal class clauses can be present, past, or future. It can have the simple, progressive or perfective aspect. The tense that is most typically found in news reports, however, is past tense, followed by present tense, with future tense occurring least frequently. This is a normal characteristic of news reports about recent or current situations. Present tense is usually used to stress the continuation or recentness of a situation. We often found the usage of present tense with the perfective or progressive aspect. Future tense is usually used in a planning context, or for a scheduled event, such as the expected time of laboratory confirmation.

4.4.2 Quantitative analysis of zone class annotation

To evaluate the reproducibility of the proposed scheme in term of event class annotation, the kappa coefficient [122] is used as agreement measurement. The following subsections report the analysis of the annotation agreement for each event class, as well as an intensive analysis of cases of disagreement.

4.4.2.1 Zone type annotation results

Table 4-2 lists the proportions of events that were classified by each annotator. The trend in classification was the same for each of the three annotators, for both corpus sets. As might be expected, the number of normal events was the highest, since news reports generally talk about current situations, which are considered normal events according to the spatiotemporal zoning scheme. The second most frequent event class was the reporting event, which is usually found in the context of reported speech, followed by events in the information and hypothetical classes.

Table 4-2: Proportions of events classified by each annotator

Corpus	Annotator	Normal (%)	Reporting (%)	Hypothetical (%)	Information (%)
Set1	A	53.31	23.30	5.80	17.59
	B	54.05	24.31	5.80	15.84
Set2	A	50.68	26.75	4.17	18.40
	C	49.89	27.53	5.51	17.07

4.4.2.2 Zone type annotation: agreement analysis

For the zone type annotation task, the kappa coefficient is used as a measure for agreement between annotators.

The results showed that the annotation scheme for zone types is reproducible, with $\boldsymbol{K}$=0.87 for annotators A and B, and $\boldsymbol{K}$=0.9 for annotators A and C. These numbers show that, given the annotation guideline, trained annotators could make distinctions between the different types of events in the same way.

In order to see which category distinctions are hard to make, I applied Krippendorff's diagnostic for category distinctions. In this setting, all other categories but the one of interest are collapsed. The results are listed in Table 4-3. Data on the number of events

was shown in Table 4-1. Krippendorff's diagnostic indicated that the most difficult distinction was one that results in best K values if omitted. From the results, the most difficult classification is that among the normal, hypothetical, and information classes. The statistics support the data observation, in that it was rarely to find reporting verbs used in the context of the information or hypothetical class, while there is confusion in categorization among the normal, information, and hypothetical classes. The lowest K value occurred for the case where hypothetical class was separated from the rest of the categories. One of the reasons that contributed to this result is a skewed number of hypothetical events in the experimented corpus. Moreover, it was found that the annotators usually have disagreement in classifying between hypothetical events and normal future events or information.

Table 4-3: Krippendorff's diagnostics for category distinction

Zone class		Annotators A and B	Annotators A and C
Normal	Reporting + Hypothetical + Information	0.85	0.91
Reporting	Normal + Hypothetical +Information	0.94	0.97
Hypothetical	Normal + Reporting + Information	0.79	0.79
Information	Normal + Reporting + Hypothetical	0.84	0.84

Another tool for analysis of annotation is the confusion matrix. Table 4-4 shows the confusion matrices between each of the two pairs of annotators: A and B, and A and C. The cells along the diagonal show the decisions on which they agreed, while all other cells show decisions on which they disagreed.

Table 4-4: Confusion matrix between annotators A and B on Set1 and between annotators A and C on Set2

		Annotator A				Total
		Normal	Reporting	Hypothetical	Information	
Annotator B	Normal	543	6	11	27	587
	Reporting	17	247	0	0	264
	Hypothetical	6	0	51	6	63
	Information	13	0	1	158	172
Total		579	253	63	191	1086

Annotator C	Normal	436	3	2	12	453
	Reporting	8	242	0	0	250
	Hypothetical	1	0	36	13	50
	Information	15	0	2	138	155
Total		460	245	40	163	908

From the confusion matrices, it can be seen that there was no disagreement in judging between the reporting class and the hypothetical or information classes, between both annotators A and B, and annotators A and C. The disagreements between annotators A and B and between annotators A and C were found mostly in classification between the normal and information classes (40 times for annotators A and B, and 27 times for annotators A and C). The high number of confusion elements between information and normal class indicates the hardness in classifying between the two categories. It signals that in classifying between normal and information class, it might be necessary to have a deeper knowledge than a surface, textual signal.

4.4.2.3 Zone type annotation: Error analysis

To gain insight into the disagreements in event classification, a detail analysis of these errors is necessary. The results of the analysis are described in this subsection.

Disagreements between the normal and reporting classes

There are certain verbs that usually cause disagreements between annotators. While there is a certain set of verbs that are always considered to indicate reporting events, such as “say”, “inform”, and “report”, there are also many verbs that can be considered to indicate either reporting or normal events, depending on the context. These verbs include “show”, “concede”, “order”, “urge”, “recommend”, “ask”, and so forth.

Disagreements between the normal and information classes

Disagreements between the normal and information classes are the most common among all disagreements. The causes of these disagreements come mainly from the following two situations:

1) Difference in perception between an event situation and a non-event situation

According to the spatiotemporal zone annotation guideline, verbs that indicate non-event situations, as in a sentence that describes the attributes or complement of its subject, are considered to indicate the information class. For example, "The victim is a 12-year-old boy" and "His condition is very severe". It was often found, however, that there were many disagreements when the attributes of the subject were represented by adjectives. An example of such a sentence is given below.

(1) A red rash **is also visible** on the bodies of affected persons.

Annotators who consider the above example as an event regard this sentence to express a perception of state by the author. This reading can be paraphrased as "I see a red rash on the bodies of affected person". I think that this type of sentence is naturally ambiguous as to whether it represents a state or an event.

2) Difference in perception of general knowledge or between a generic event and a specific event

According to the spatiotemporal annotation guideline, a typical case of the normal class is temporally located and described by clauses with specific subjects, while a typical case of the information class is not temporally located, and the subject of the sentence can refer to either specific or non-specific entities.

Generally, when the verb's subject is considered to refer to non-specific entities, this event should be classified as the information type. Different annotators, however, might have different views of the verb's subject, with one annotator considering the subject to refer to a general entity and another considering it to refer to a specific entity. The classification becomes more complicated when a temporal expression appears in the sentence, as in the following examples. When there is such a temporal expression, one might feel that the event can be located in time and thus classify it as the normal type. This situation, as shown in the examples, causes disagreement between annotators.

(1) People working in the wool industry **used to be prone** 50 years ago.

(2) The mushroom **has been eaten** in Japan for centuries.

These two situations happened quite often, and their judgment depends on the experience and perspective of each annotator. A decision tree that provides an explicit guideline to annotators in annotating the zone type might help reduce this type of disagreement

Disagreements between the normal and hypothetical classes

Disagreements in this group come mainly from different interpretations between normal future events and hypothetical events. This situation occurred when one annotator felt that the linguistic element of focus (e.g., a sentence) represented something that will definitely occur in the future (i.e., a normal event), while another annotator thought that the element referred to a prediction or a conditionally possible situation (i.e., a hypothetical event). The example below illustrates this disagreement:

> "Our hospitals are not clean and we **will find** when this infection has passed that we **will have** another that takes its place. Until we **get** our hospitals clean we**'re going to have** infections all the time."

In the above example, while one annotator annotated the events marked in boldface as normal events, the other annotator considered them as hypothetical events, giving a prediction of a situation.

According to the data analysis, there were disagreements in deciding whether "would" was used to signal the future aspect or the hypothetical sense (i.e., indicating possibility or willingness), as in the following example:

> The Red Cross said it **would spend** nearly one million Swiss francs (602,000 euros / 867,000 dollars) in a four-month awareness drive.

This situation occurred mostly in indirect speech construction. In conversion of direct speech to indirect speech, the tense of verbal elements within the reported speech is changed according to tense of the reporting verb. When the reporting verb is in past tense, such as "said", the modal "will" is usually changed to "would", while "would" is left as is. This construction sometimes causes difficulty in judging an original form of "would" used in indirect speech. Since the auxiliary "*would*" usually signals the conditional or optative mood (e.g., the hypothetical type), while "will" signals the future aspect of an event occurring, different judgments may result in different classification of such events.

Disagreements between the hypothetical and information classes

Disagreement in terms of the hypothetical and information classes occurred very often when there was hypothetical mention of general concepts or general knowledge, as in the following example:

Because West Nile virus antibodies can stay within a person's bloodstream for up to 500 days, it **can be difficult** to determine the date of infection.

While one annotator viewed "can be difficult" as indicating generic information about West Nile virus, the other annotator considered it to indicate a hypothetical situation relating to a certain West Nile virus infection.

4.4.3 Zone spatial attribute annotation results

Since annotating the location attributes of marked-up events is not exactly a classification task, the percentage agreement is used as a measure to obtain inter-annotator agreement statistics.

4.4.3.1 Spatial attribute annotation: Agreement analysis

The agreement statistics on the location attributes of marked-up events are listed in Table 4-5. The numbers in the table were calculated only from pairs of events that were classified into the same class by both annotators.

Table 4-5: Agreement statistics (percent agreement) for location attribute annotation

Annotators		Normal	Reporting	Hypothetical	Information	All classes
A and B	Strict	0.82	0.84	0.80	0.70	0.806
	Loose	0.99	1	1	1	0.997
A and C	Strict	0.75	0.78	0.58	0.72	0.749
	Loose	0.99	0.98	0.89	1	0.986

With strict analysis, only location attributes that were annotated with exactly the same location(s) would be considered to indicate agreement in annotation. According to the annotation results, the annotators seemed to disagree on location selection more often for events in the hypothetical and information classes than for events in the normal and reporting classes. For the information class, disagreements occurred most often when the event to be annotated consisted of general knowledge. I hypothesize that it is more natural for a human to consistently locate events that actually occur than to specify the locations of non-occurring events, such as information, as in the following example:

Human T-cell Lymphotropic Virus, Type1 (HTLV-1) **occurs** mostly in Japan, Caribbean countries and Africa. Doctors say most people who contract it **will show** no symptoms, but in

about five percent of cases, it **can lead** to cancers of the blood and diseases affecting the nervous system.

The events marked in boldface were classified as the information type by both annotators. While annotator A considered these events as world knowledge, selecting the location as "World" for all three, annotator C considered them as information about specific locations and selected Japan, Caribbean countries, and Africa as the locations of these information events.

After thoroughly examined the data, it has been found that even when the annotators selected different locations, these locations mostly appeared to be related. Specifically, either the locations selected by one annotator are located within the location(s) selected by the other annotator, or the locations selected by both annotators are partially the same. Especially in the case of locations that are partially the same, I observed that many of these selections occurred when one annotator selected only locations at a lower level of administration (such as selecting only villages), while the other annotator selected locations at both lower and higher levels of administration, which included the lower-level locations. Although these annotations do not represent 100% agreement, they are not totally different. They simply convey different levels of information. With loose agreement analysis, in which partial agreement or inclusion of a location is acceptable, the percentage agreement was very high, at almost 100% for most event classes for annotators A and B. The situation was the same for annotators A and C, except for the hypothetical class, in which the agreement was a little bit lower.

For spatial attribute annotation, the inter-annotator agreement was also evaluated based on one another parameter, which is Krippendorff's α [124, 125]. The Krippendorff's α reports the degree that the observed number of agreements could be expected to occur by chance. The value of α range from 1 to -1, with 0 representing that the agreements observed could completely occurred by chance. Based on the analysis of the annotation results (strictly analysis on overall events), the agreement between annotators A and B was $\alpha = 0.76$, while the agreement between annotators A and C was $\alpha = 0.70$.

In surveillance systems, alerts currently are geo-coded at the country scale, with province-, state-, or city-level resolution for select countries. Viewing the percentage agreement as an approximate maximum score for automatic annotation, the results show that reliable, automatic event location recognition can be obtained at the country level.

According to the further analysis of the data, I noticed that spatial annotations of normal events usually had agreement or partial agreement at the state or province level, especially in well-characterized countries. This result indicates a promising possibility for identifying outbreak locations with a finer geographic resolution, which is a critical area in future development of effective outbreak detection.

Note that the annotations were done without providing the annotators access to a gazetteer or geographical ontology. I believe that with full access to geographical data, the agreement between annotators should be improved.

4.4.3.2 Spatial attribute annotation: Error analysis

According to the examination of the raw data to find the characteristics of disagreements between annotators, I observed that the disagreements mostly occurred for the following reasons.

1) Event related to the movement of an object

There are certain sets of event-representing verbs that convey a sense of spatial movement, i.e., verbs that can be used with a prepositional phrase like “from <place>” and “to <place>”. These verbs include “transfer”, “send”, “travel”, “move”, and so forth.

It was often found that either the source or destination location was mentioned explicitly, but not both. When only the source or destination location was stated, partial disagreements between annotators usually occurred. Disagreement occurred when one annotator selected only a source (or destination) location that appeared nearby in the same sentence, while another annotator also attempted to infer the destination (or source) location from the discourse. Disagreement also occurred when both annotators attempted to infer missing locations, but the results of inference were not the same.

2) Location information via inference

Locations to be selected as event locations may be stated directly or can be inferred from context or discourse. When they were stated directly, the annotators were capable of choosing event locations with 100% agreement. When event locations were not mentioned explicitly, however, disagreement between annotators could occur. Without such explicit information at hand, I often found that while one annotator tried to infer the most specific locations according to what was available in the news content, another annotator tended to select locations with at a higher level of administration, such as a

location at the country or province level, whenever there was uncertainty. The following are examples of these situations:

> Dr. Ruth Improso **denied** that the number of typhoid fever cases in Bunawan town, particularly in Barangay (village) Libertad, went up to as high as 500.

For reporting events, annotators generally selected the location of the reporting agency as the event location. When the annotators did not have this information, however, there could be disagreement. In the above example, one annotator selected the fine-grained location reported to have experienced the outbreak, which was Bunawan. Another annotator, however, concluded that there was not enough information to assume that the "denied" event occurred in Bunawan and selected the Philippines as the event location, which is more general.

> Mekong Delta provinces are in the grip of a dengue outbreak with 38% more patients year on year. Measles is also afoot in northern Lai Chau Province. Deputy Minister of Health Trinh Quan Huan announced news of the outbreaks recently, saying that measures **were underway** to prevent further spread.

In the above example, while one annotator selected Mekong Delta provinces and Lai Chau as the locations of the "were underway" event, another annotator doubted whether the measures were underway only in these affected provinces, and decided to select Vietnam, which is more general, instead.

> Iam Tanthong, a patient from Ban Paa Ngiw in Wiang Pa Pao district, said she had been **admitted** to hospital complaining of serious abdominal pain.

When the location of a hospital was unknown to the annotators, it could lead to inconsistent annotation. I found that while one annotator selected the hometown of the patient as the location of the "admitted" event, another one decided to select the most certain location, which was the city or country where the situation had occurred.

3) Disagreement in interpreting the location of an event

This kind of situation did not occur very often, but the annotators could sometimes be misled by unclear passages, as in the following example:

> So far, there's no hint of an outbreak in Canada. But Canadian health officials are watching what happens in the U.S. They may just **start testing** birds here to find out if they're carrying the virus. Because if they've **got it**, mosquitoes **will pick it up,** and then, people **will be** next.

While one annotator considered the events “start testing”, “will pick up”, and “will be” related to a hypothetical situation in Canada, another annotator chose the U. S. as the event location.

4.4.4 Zone temporal attribute annotation results

The temporal attribute annotation task uses the same measurement calculation as that for location agreement analysis, i.e., the percentage agreement.

4.4.4.1 Temporal attribute annotation: Agreement analysis

In agreement analysis for temporal attribute annotation, I considered the temporal attribute to be the same only when all of the temporally related attributes of each event were consistently marked up by the two annotators. The agreement statistics for temporal attributes are listed in Table 4-6.

Table 4-6: Agreement statistics for temporal attributes annotation

Annotators		Normal	Reporting	All classes
A and B	Strict	0.92	0.97	0.94
	Loose	0.98	0.99	0.98
A and C	Strict	0.95	0.89	0.93
	Loose	0.99	0.95	0.97

For strict analysis, only events that were marked up with exactly the same set of temporal attributes were counted as agreement in annotation. In loose analysis, I considered any pair of annotations with the same VAL attribute as indicating agreement.

From the results, the agreement on temporal attributes was very promising for both pairs of annotators, even with strict analysis. This indicates that temporal annotation was less confusing for human annotators than location annotation, and that the instructions for temporal annotation were reproducible.

4.4.4.2 Temporal attribute annotation: Error analysis

In order to locate the causes of disagreement, the annotated documents were thoroughly investigated. It has been observed that disagreements mostly occurred when temporal information was not stated directly but could be inferred from the discourse.

There is an idiosyncrasy in news reports in which the first paragraph is used as a summary of the news story. In the case of an interview with a person in charge, the time of the interview is usually given in this first paragraph and then omitted in the rest of the story. Moreover, in this paragraph, interviewed people are usually referred to by a short description, such as "doctors from several hospitals" or "senior health officials". This usually caused disagreement between annotators, especially in long articles. Each annotator might have judged differently whether each person appearing in the story was the same person or was part of a group mentioned in the first paragraph. This led to inconsistency between annotators in selecting temporal attributes. Figure 4-6 shows a capture of text representing one example of this situation.

The patients suffering botulism poisoning who were rushed to Bangkok from Nan remain in a critical condition, doctors at several hospitals ***said*** **yesterday**.
While the 17 people are stable, they are not yet out of danger, Medical Service director-general Dr Chatri Banchuen ***said***.
...
..., Chatri ***added***.
...
..., hospital director Dr. Jessa Chokedumrongsuk **said**.
...
..., hospital director Dr. Vinit Pua-pradit ***said***.
...
..., the doctor ***claimed***.

Figure 4-6: Example of co-referring of events

Text capture in figure 4-6 exemplifies the situation which multiple clauses refer to the same real-world event. In the text example, the phrase "Medical Service director-general Dr Chatri Banchuen said", "Chatri added", "hospital director Dr. Jessa Chokedumrongsuk said", "hospital director Dr. Vinit Pua-pradit said", and "the doctor claimed" are parts of the event previously mentioned in the clause "doctors at several hospitals said yesterday".

Gerunds and infinitives were another source of disagreement in temporal annotation. The surface forms of these expressions are tenseless. Without tense, which is a basic signal of temporal information, annotators sometimes had different opinion about an event's time.

Another situation which was found quite often is the disagreements occurred when there was a temporal expression in a relative clause, as in the following example:

It **had reports** of 39 deaths from the outbreak of suspected acute hemorrhagic fever which began in January.

There could be disagreement in that one annotator felt that the "had reports" event occurred in the same period as the beginning of the outbreak, while another annotator thought that the "had reports" event could have occurred at any time after the beginning of the outbreak.

Different judgment of the time span or length of an event was another cause of disagreement, as in the example below:

On Christmas day, a 24-year-old woman from Jakarta also **died** from the virus after **buying** a live chicken from a market.

In the above example, while annotator A viewed "buying" as an event that occurred before Christmas day, annotator B considered both "died" and "buying" to have occurred on the same day, i.e., Christmas day.

Chapter 5

Automatic Approaches for Spatiotemporal Zoning Annotation

This section describes the strategies for automatic spatiotemporal zone annotation. The methodology for the automatic zone annotation was proposed and evaluated for each group of zone attributes, which are event type recognition, temporal attributes recognition, and spatial attribute recognition.

To automatically classify event expressions, i.e. zone type recognition, machine learning technique was employed for incorporating various sets of textual features. To recognize spatial information, several approaches, ranging from simple technique such as commonly used heuristic-based approach to a more sophisticated machine learning approach were studied. Various textual features and the strategy for feature encoding were explored in order to effectively recognize spatial attribute of the events. For temporal attribute recognition, traditional rule-based approach was used to recognize an event's temporal information.

The evaluation of the approach for automatic zone attribute annotation employed an n-fold cross validation strategy. The experimental results and analysis are also described in detail in this chapter.

5.1 Zone generation process

The task of spatiotemporal zoning can be separated into 3 main steps. (1) Document pre-processing: location names, temporal expressions, and clause boundary in the documents are identified and marked-up. This provides the basic elements for zone attribute analysis and can be done automatically using natural language processing software [45, 120, 121]. (2) Attribution 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. If the consecutive event-predicates have the same attribute values, they will be merged into a larger zone unit. Otherwise, they will

be marked as different zones. To provide further insight into the zone boundary generation task, the process of boundary generation is illustrated in the figure below.

Suppose: ep_i = event-predicate *i* that does NOT appear inside the sub-ordinate clause
ep_i' = event-predicate *i* that appear inside the sub-ordinate clause
$subj_i$ = subject of ep_i (including all phrases that modify $subj_i$)
obj_i = object of ep_i (including all phrases that modify obj_i)
mod_i = ep_i's modifiers
$subor_i$ = sub-ordinate clause of ep_i

```
Zone_Boundary_Generation
{       Scan document for the first event-predicate ep1
        BoundaryExtension(document, 1, ep1)
}
Boundary_Extension
{   Input: Text, i, epi,
    (1) Check for the epi's property
          if i=1 or epi's attributes are not the same as epi-1
             Initialize a new zone with the attributes
             of epi's attributes.
    (2) Extend the zone boundary from epi to the left
        and right to cover all subji, obji, modi, and
        subori that appear before the occurrence of
        epi+1.
    (3) If the boundary cover sub-ordinate clauses
          For each sub-ordinate clause, subori
             Scan subori for the first appearance of ep1'
             BoundaryExtension(subori, ep1', 1)
    (4) Scan for the next event-predicate, epi+1, in Text
        (4.1) Update value of epi to epi+1
        (4.2) Update value of i to i+1
    (5) Repeat step (1)
}
```

Example

[The health officials in Pakistan *have confirmed* that
    [the Crimean-Congo hemorrhagic fever *has killed* at least 5people , including a woman doctor , and *infected* around 45 people in the country's biggest city of Karachi.]
]
[Dengue, commonly known as Congo fever, *is spread* through the bite of an infected Aedes Aegypti mosquito.]

Figure 5-1: Zone generation process

The boundary marked with the square brackets in the text capture illustrated in the Figure 5-1 is the example of the output from the zone boundary generation process. As shown in the figure, the text was annotated as follows. Start at event-predicate "have confirmed", the boundary of the first zone will be extended to cover the subject (The health officials in Pakistan) and the sub-ordinate clause (the Crimean-Congo …) of the event-predicate, and then move to the second event-predicate "is spread". Since the class of the second event-predicate is "Information", which is different from the first event-predicate, it is marked up in a new zone. The next step is to analyze the event-predicates inside the sub-ordinate clause. The attributes of "has killed" and "infected" are compatible to each other, so they are marked in the same zone.

Since the zone boundary generation task (3) is relatively trivial when all attributes are known, this thesis focuses here on the study and evaluation of attribute annotation (2).

5.2 Automatic approach for event type annotation

To recognize the class of text-based event, I formulated the task as classification problem and employed the statistical machine learning technique, namely Conditional Random Fields (CRF). In the experiment, various combinations of linguistic features were incorporated to test which feature set will give the best performance.

5.2.1 Previous approach for text content classification

The previous works of automatic text content classification were proposed in [118] for classifying area of text to (possibly overlapping) rhetorical zone classes according to argument and intellectual property of text. These works employed supervised machine learning approach in recognizing class of scientific text content. Tuefel [96, 118] proposed a strategy to classify text into 7 classes by employing machine learning technique, such as Naïve Bayes model, N-gram models. The Features used in this works concerned about, such as occurrences of certain words in the sentence, position of the sentence, length of the sentence, syntactic information of verb phrase, appearances of citation in the sentence, appearances of formulaic expression, type of agent, type of action. In [126], the different features for learning rhetorical zones were explored and use Naive Bayes and Support Vector Machine (SVM) for zone classification. Features used in this works came from lexical/syntactic information, main verb, location within text, and zone sequence.

The problem of event type annotation in this thesis is resembled the same characteristic as those works mentioned previously in that it can be viewed as a classification problem. More specifically, the event type annotation is the task that attempts to classify text content regarding to the spatial and temporal characteristics. However, the nature of classes for events to be classified is different from the rhetorical or argumentative zoning, so some features used in these works may not suit to this task. In the following subsections, the approach for event type annotation and the features that suit for the task are explored and discussed.

5.2.2 Event type annotation approach

5.2.2.1 Conditional Random Fields

There are many machine learning algorithms which are able to classify items into predefined categories, given a set of sentential features. Supervised learning methods take the correct answer of the classification into account which must be provided externally whereas unsupervised techniques learn without such external provision of the correct answer. To solve the problem of automatic event classification, one supervised machine learning technique called Conditional Random Fields (CRF) [127] is employed. In this subsection, a brief detail of this technique is introduced.

Conditional random fields (CRF) is a probabilistic framework for labeling and segmenting sequential data. A CRF is a form of undirected, graphical model that defines a single log-linear distribution over labeled sequences for a particular observation sequence [128]. CRF is a rich context/featured based tool of classification, where the context/features may be redundant, overlapping and pre-classified. The overlapping and redundant features mean that one can extract both specific and generic information to be used in classification. Pre-classified feature means that each feature affects different classes differently. This kind of classification model is good for training both for small and large corpora, compared to a joint probability model such as naive Bayes or Hidden Markov Model (HMM). The main differences between this and other conditional probability models is that its output is structured, and compared to Maximum Entropy and Maximum Entropy Markov Model (MEMM) it allows to avoid the label bias problem.

The CRF used in this work is a conditional probability classifier that segments and labels sequential data as described in [127]. CRF define conditional probability distributions $p(Y|X)$, where Y is a label sequence and X is an input sequence where a linear-chain CRF is a distribution $p(y|x)$ defined as follows.

$$p(y \mid x) = \frac{1}{Z(x)} \exp\left\{\sum_{k=1}^{K} \lambda_k f_k(y', y, x)\right\} \qquad (1)$$

where; λ is the weight of the k^{th} feature

$f_k(y', y, x)$ is the k^{th} feature from a feature set of size K.

y' is the previous label or state

y is the current label or state

x is the current input

$Z(x)$ is the normalization function defined as

$$Z(x) = \sum_{y} \exp\left\{\sum_{k=1}^{K} \lambda_k f_k(y', y, x)\right\} \quad (2)$$

The most probable label sequence for the input sequence X is

$$Y^* = \arg\max_{y}(p(Y \mid X)) \quad (3)$$

5.2.2.2 Information source for event classification

In the zone type annotation task, I found that the occurring of some set of words, such as modal verb, adverb of frequency, in verb phrase usually has impact in classifying event. Modal verb usually indicate information or hypothetical class, while the presence of adverb of frequency usually indicate information class. Reporting event generally represented by a certain set of verbs, such as "say", "tell", "report", etc. Moreover, when the verb phrase is composed of verb to be and adjective, it is usually classified as information class. However, there are some adjectives that are considered as representing event in normal class, such as "ongoing", "underway", which relate to process, or adjectives that express feeling, such as "happy", etc. From this observation, set of word lists was created according to their special characteristic to the event classification task. These word lists will be use as features in the learning model. The features used in event classification task are shown in the Table 5-1.

Table 5-1: Linguistic features for event classification

Feature	Description
Internal verb phrase features (1) Verb stem	Stem form of main verb Example: 'have died' -> 'die' This feature includes checking for 1) Check whether verb phrase consists of verb to be and adjective, such as 'is big', etc. 2) Check whether verb phrase consists of verb to be and prepositional phrase, such as 'is in the hospital', etc.

	(2) Verb tense	Tense of verb Example: ‘have died’ -> ‘Present perfect’
	(3) Contain adverb of frequency	Verb phrase contain adverb of frequency, such as ‘always’, ‘usually’, etc.
	(4) Contain modal	Whether the verb phrase contain modal verb Modal set I: such as ‘can’, ‘should’, ‘must’, etc. Modal set II: such as ‘may’, ‘might’, ‘would’, etc.
	(5) Is_Report	Check stem form of the verb phase whether it is a verb with communication function, such as ‘say’, ‘report’, ‘announce’, etc.
	(6) Contain adjective	Whether the verb phrase contain adjective in the list, such as ‘ongoing’, ‘undergo’, etc.
	(7) Contain future modal verb	Whether the verb phrase contain modal verb, such as “will”, “shall”
Contextual features	(8) Hypothetical clue	Check whether: - In-focus verb phrase is in the subordination clause of verb phrase that contain hypothetical signaling word, such as ‘suspect’, ‘fear’, etc. - There is a hypothetical signaling word, such as ‘possible’, around in-focus verb phrase.
	(9) Is subject general	Whether subject of the in-focus verb phrase refers to non-specific entity
	(10) Subject type	Category of subject, which are: “Disease_Germ” (i.e. disease or pathogen) “Symptom” (i.e. symptom of any disease) “Officer” (i.e. person who has the property of being an official, such as government or medical officer) “Civilian” (i.e. person who does not have the property of being an official) “Gov. Organization” (i.e. government organization) “WHO” (i.e. World Health Organization) “WHO-related” (i.e. WHO agency or UN agency) “Organization” (i.e. other kinds of organization) “Location” “Other” (none of the above class)
	(11) *n*-left words	*n*-context words on the left of in-focus verb, here I use $n = 3$.
	(12) *n*-right words	*n*-context words on the right of in-focus verb, here I use $n = 3$.
	(13) Previous event category	Class of previous verb phrase
	(14) In if-clause	Check whether in-focus verb phrase occurs in if-clause

5.2.3 Event classification Results

In the experiment for event classification, one statistical machine learning model, namely Conditional Random Fields, was employed. Various combinations of the features mentioned in the previous section (cf. Table 5-1) were incorporated into the model. In order to study the contribution of each feature, the two other sets of experiments were also conducted. In these experiments, each feature was removed from the model in an excluding-one-feature-per-training manner. The event classification results are shown in the Table 5-2.

Table 5-2: Evaluation of the contribution of contextual features in event classification [3]

Features	Class (%F-score (%Precision, %Recall))				Overall (%Precision)
	Normal	Reporting	Information	Hypothetical	
All features (Baseline)	90.9 (89.7, 92.0)	96.4 (94.9, 97.8)	64.2 (66.2, 62.4)	51.5 (63.0, 43.6)	86.9
Without hypothetical clue	91.0 (89.8, 92.2)	96.4 (94.9, 97.8)	65.0 (66.9, 63.1)	46.2 (57.7, 38.5)	86.9
Without Subject type	90.0 (88.8, 91.2)	95.5 (94.1, 97.0)	62.1 (63.2, 61.0)	41.9 (56.5, 33.3)	85.6
Without checking whether subject is general	91.2 (90.7, 91.6)	96.0 (93.8, 98.3)	66.9 (67.1, 66.7)	38.1 (50.0, 30.8)	86.9
Without n-left words	90.7 (90.2, 91.2)	96.1 (94.9, 97.4)	64.8 (65.0, 64.5)	50.7 (60.7, 43.6)	86.7
Without n-right words	**92.7** (92.4, 93.0)	95.9 (94.9, 97.0)	**73.9** (73.4, 74.5)	52.9 (62.1, 46.2)	**89.2**
Without previously predicted event class	91.0 (89.9, 92.2)	96.4 (94.9, 97.8)	65.7 (67.7, 63.8)	45.5 (55.6, 38.5)	87.0
Without checking whether the event is in if-clause	90.1 (88.8, 91.4)	95.9 (94.9, 97.0)	60.8 (62.9, 58.9)	47.8 (57.1, 41.0)	85.7
Without contextual features	87.6 (94.4, 81.7)	**96.6** (94.6, 98.7)	65.9 (55., 80.9)	**57.5** (61.8, 53.8)	84.6

In Table 5-2, the contribution of contextual features was evaluated. According to the results, compared to the model trained with all the features, removing features from *n*-right context words improved the performance of the classification model (from 86.9% to 89.2% precision). This indicated that incorporating *n*-right context words has negative impact on the classification. Without subject type feature, the classification performance was the lowest (from 86.9% to 85.6% precision).

[3] Features mentioned in this table are described in detail in table 5-1

The performance was also analyzed for each event class. For normal event, the feature that has the most negative impact on event classification (i.e. causes the improvement of performance when excluding the feature from the model that incorporates all contextual features) is n-right context words (from 90.9% to 92.7% F-Score). For reporting event, the results show that the performance was reduced the most when the subject type feature was removed from the model (from 96.4% to 95.5% F-score). For hypothetical event, excluding the feature that indicates whether the subject refers to specific or generic entity causes the significant drop in performance (51.5% to 38.1% F-score). While most of the contextual features seem to be useful for recognizing hypothetical event, removing *n*-right context words, on the other hand, caused the improvement of the performance (raised to 52.9% F-score). For information class, removing subject type feature and in-clause feature has negative impact on classification performance (from F=64.2% to 62.1% and 60.8% respectively). Excluding *n*-right context words seems to raise the classification performance significantly (F=73.9%).

Table 5-3: Evaluation of the contribution of internal features in event classification [4]

Features	Class (%F-score (%Precision, %Recall))				Overall (%Precision)
	Normal	Reporting	Information	Hypothetical	
All features except right context (Baseline)	**92.7** (92.4, 93.0)	95.9 (94.9, 97.0)	**73.9** (73.4, 74.5)	**52.9** (62.1, 46.2)	**89.2**
Without Adjective feature	91.9 (91.2, 92.6)	95.9 (94.9, 97.0)	68.8 (69.6, 68.1)	50.0 (58.6, 43.6)	87.9
Without Future modal feature	92.7 (93.4, 92.1)	95.9 (95.3, 96.5)	71.9 (72.1, 71.6)	46.4 (53.3, 41.0)	88.6
Without checking whether it is Reporting verb	90.8 (89.3, 92.4)	91.2 (92.8, 89.6)	72.3 (72.3, 72.3)	52.9 (62.1, 46.2)	86.7
Without Verb stem feature	90.4 (90.0, 90.8)	95.2 (94.8, 95.7)	66.7 (66.0, 67.4)	47.1 (55.2, 41.0)	86.3
Without Adverb of frequency feature	92.6 (92.6, 92.6)	95.9 (94.9, 97.0)	73.4 (72.4, 74.5)	52.2 (60.0, 46.2)	88.9
Without Modal feature	90.6 (90.5, 90.6)	**96.1** (94.9, 97.4)	69.7 (68.5, 70.9)	42.4 (51.9, 35.9	87.0
Without Tense feature	90.8 (90.0, 91.6)	95.9 (95.3, 96.5)	68.8 (69.6, 68.1)	52.2 (60.0, 46.2)	87.3
Without internal features	84.5 (81.3, 87.9)	91.3 (93.6, 89.1)	51.5 (53.4, 49.6)	24.6 (38.9, 17.9)	79.4

[4] Features mentioned in this table are described in detail in table 5-1

In Table 5-3, the contribution of internal features was evaluated. In this experiment, the base model for comparison is the model that incorporates all internal and contextual features, except n-right context words. This model was selected as the base model since it performed the best in the experiment that evaluated the contribution of contextual features. According to the results, comparing to the base model, removing the verb stem feature caused the most drop in performance (reduced to 86.3% precision). In the analysis of each event class, for normal and information class, removing the verb stem feature has the most negative impact; reducing the performance from 92.7% to 90.4% F-score for normal class and from 73.9% to 66.7% F-score for information class. The feature that checks for reporting verb has the most impact on the classification of reporting class. Removing this feature caused the most reduction in performance (from 95.9% to 91.2% F-score). On the other hand, the modal feature seems to have negative impact to the reporting event recognition. Removing this feature improves the performance to 96.1% F-score. For hypothetical event, the results show that the performance was reduced the most when the modal feature was removed from the model. The results also indicated that the internal features are more important than contextual features. Without utilizing internal features, the performance was reduced significantly (794% precision).

Results analysis and discussion

In order to get insight into the errors occurred in the classification, the results from the best model, the one that was trained with all features except right context, were analyzed in detail. Table 5-4 illustrated the proportion of the events that were miss-classified by the model. In the table, the correct classification of the events are represented in the first column, the classes that were miss-classified by the model are represented in the first row. The interpretation of the Table 5-4 is, for example, based on the overall errors, the proportion of normal events that were incorrectly classified as information is 31.37%.

Table 5-4: Proportion of events that were miss-classified by the model

Event class	Info	Normal	Report	Hypo
Info		30 (27.27%)	2 (1.82%)	7 (6.36%)
Normal	30 (27.27%)		9 (8.18%)	2 (1.82%)
Report	1 (0.91%)	6 (5.45%)		1 (0.91%)
Hypo	11 (10.00%)	8 (7.27%)	3 (2.73%)	

The statistics illustrated in the table showed that the errors in classification were mainly found in classifying between normal and information class. This result is the same as the result from human annotation where the disagreements were found mostly between normal and information class. In order to improve the precision in differentiating between normal and information class, encoding of deeper knowledge may be necessary. The confusion between hypothetical class and information class is also obvious. This could contribute to the overlapping between hypothetical situations embedded in the information of general things. For example, the text excerpt illustrated below describing general fact about the disease. The general fact also included possibilities for certain circumstances (in the example below, "may lead to vomiting"), which imitated the characteristics of hypothetical events.

> Pertussis begins with cold symptoms as sneezing, a runny nose, a low-grade fever, but the cough becomes more violent and *may lead to vomiting*.

This situation become more problematic where there is only a small number of events in hypothetical class in the training corpus, which could cause a low statistic in the learning process.

5.3 Automatic approach for spatial attribute annotation

The previous works in processing text-based spatial information usually in the form of location entities recognition [43-45, 129], or geographical grounding system [67-69]. The geographical grounding systems aim to associate textual location expression with the real world geographical location, which can be, such as geographical ontology, or geographic coordinates. Although these works provide critical information in solving location ambiguity and precisely visualizing geographical data, they do not associate these locations to events or actions mentioned in text. Nevertheless, the ability to anchor events with the locations in which they occurred would assist for an effective analysis of situations being reported in articles. In this section, the strategies for associating events with their spatial locations were proposed and evaluated.

5.3.1 Information source for spatial attribute recognition

In order to develop an effective automatic system for recognizing a textual event's spatial attribute, the features to be used as the information source for the recognition task must be carefully selected. I asked 10 people, including the 2 linguists, 7 Ph.D. students in the Department of Informatics at The Graduate University for Advanced Studies, and 1 pose-doctoral researcher, about how they recognized the place where the event reported in the news occurred in order to gather their opinion. The results from the observation showed that they usually agreed on the following textual information sources to be used for identifying the place of occurrence of the events. The source are listed below. In the following description of textual features, a "verb" means a verb phrase that represents the action or state. A "subject" means the subject of the verbal expression representing the action or state. An "object" refers to the direct or indirect object of the verbal expression representing the action or state.

- Location of subject

This feature represents the (possible) geographical location of the subject. They can be the location name in the subject modifier, the location name in the subject appositive, and the location name in the subject's relative clause among others. For example:

(1) Head of *South Halmahera* district health office, Dr Abdurrahman Yusuf **confirmed** the spread of diarrhea and malaria in the villages

(2) Last week's floods which hit *Songa* and *Silang* villages **had killed** hundreds of people.

- Location of subject's co-reference

This feature refers to the location names that appear in or modify the noun phrase that corefers to the same real-world entity as the subject of the in-focus verb. For example:

(3) A local health official in *Calgary* region says ....

The official **says** that ....

In the above example, "The official" co-refers to "a local health official in Calgary region".

- Location of object

This feature represents the (possible) geographical location of the object. It can be a location name in the object modifier, a location name in the object appositive, or a location name in the object's relative clause. For example:

(4) The diseases also **hit** many areas in the district capital of *Labuha*.

- Location of object's co-reference

This feature refers to the location names that appear in or modify the noun phrase that corefer to the same real-world entity as the object of the in-focus verb. For example:

(5) At least five people reportedly died of diarrhea and malaria in South Halmahera district, North Maluku province, a local official said.

.....

Flash floods **hit** some areas in *the district* several weeks.

In the above example, "The district" refers to Halmahera district.

- Location of verb

The location of the verb is considered to be location names that appear in the phrase that directly modifies verb, such as location name in prepositional phrase, etc. For example:

(6) Three people **had died** in *Makassar*.

- Location of verb's coreference

Here, verb coreference means a pair of verbs that have same meaning (or the same sense) and subjects of both verbs refer to the same real-world entity. For example:

(7) The patients suffering botulism poisoning who were rushed to Bangkok from Nan remain in a critical condition, doctors at several hospitals in Bangkok **said**$_1$ yesterday.

.....

While the 17 people are stable, they are not yet out of danger, Dr. Chatri Banchuen **said**$_2$.

.....

"Basically, their condition remains unsafe," he **said**$_3$.

In the above example, "doctors at several hospitals in Bangkok said", "Dr. Chatri Banchuen said", and "he said" all corefer to the same event. Since they all refer to the same event,

the place where $said_1$ occurred can be used as a signal for a place where $said_2$ and $said_3$ occurred.

- Inference or co-reference of locations that appear in verb modifier

In some cases, there is no obvious information about the event's geographical location appearing in the clause or sentence, such as, when the location appeared in a verb's modifier phrase is not geographical location. However, human readers can infer the geographical location from the story.

(8) The victims, who were practically of all ages and both sexes, began feeling intense pain Saturday afternoon and were rushed to three hospitals in *Magelang*, namely Tidar, Harapan and Lestari hospitals.

Up till 22.0 pm, many patients including the wedding party's host, Sudinem and the bridegroom, Dwi Purwanto, were still being treated at Tidar hospital`s emergency unit.

However, the bride, Eni Aryani , was well and **was sitting** beside her husband *in hospital*.

From discourse-level information, it can be inferred that the event "was sitting" occurred in Magelang, which is geographical location of the hospital, Tidar.

- News agency location

The location of news agency, especially local news, can sometimes be used as a default location of the situation reported in the news article.

- Nearest location names

Previously mentioned location(s) that is closest to the event expression can usually be used as a clue for recognizing the actual location of that event. In this thesis, the nearest locations were considered according to the following heuristic:

For representing nearest location names, I created two sub-features, which are the country-level nearest location and the sub country-level nearest location. Both sub-features use the same heuristic for extraction.

- Location in news headline

Under certain circumstances, the location name appearing in the headline can be used as a default location for the situation reported in the news articles

- Previous verb's location

Without introducing an expression in indicate a change of geographical location, human readers usually perceive the continually reported story as occurring in the same place. From this observation, the location recognized as a place where the event represented by textually-previous verbs occurred is introduced as another feature for spatial attribute recognition.

The above linguistic-based information sources are used as features for automatically recognizing spatial attribute of events.

5.3.2 Spatial attribute recognition: Experimentation

This subsection discusses the approaches for automatically recognizing an event's locations. For automatic spatial attribute annotation, 3 experiments were conducted in order to evaluate 3 different methodologies, which range from the simple heuristic approach to a more sophisticated statistical machine learning approach. The detail of each methodology is described next in the subsequence sections. For evaluating the approach for automatic spatial attribute annotation, I employed *n*-fold cross validation strategy and used the corpus set 1 for the experiment.

5.3.2.1 Heuristic approach: textually-closest location names

To recognize an event's locations, I first start with a heuristic-based approach, which can be considered as a baseline system. This approach relied on simple heuristic rules in identifying location of the events. The heuristics used here are: If there is a location name or a set of location names that appear in the same sentence as the event expression, then the set of location names will be considered as a location where the event expressed by the event expression occurred. Otherwise, the system will select location name(s) in a clause that appear before and was textually-closest to the event expression as a location of event.

Experimentation results

The results of the attribute annotation according to this approach are shown in the Table 5-4 below. The annotation scores were separated according to the event class in order to see the performance of this approach on different event class.

Table 5-4: Experimentation results for recognizing spatial attribute of the events based on the heuristic approach

Event type	Precision	Recall	F
Normal	65.6	59.5	62.4
Report	36.2	42.5	39.1
Information	32.4	30.1	31.2
Hypothetical	60.5	52.3	56.1
Overall	52.8	51.2	52.0

According to the results, although this approach is very simple, the results are quite low, especially for the information class. After thoroughly examining the results, I observed that the errors occurred for many reasons, for example:

1) The closest location names occurred in an adjective clause that modified the noun that was not directly related to the event.

2) The flow of the news story first discussed the main events, then moved onto the discussion about the sub-events or detailed information, which occurred in different locations from that of the main events, and finally moved back to the main events. Incorrect locations can be recognized when the report doesn't restate the location where the events in the main topic occurred. This situation causes the closest location names to become the locations of the sub-events.

3) In many cases, general knowledge, such as that about a disease or pathogen, usually involves mentioning the location. This kind of information is generally considered world knowledge, i.e., should not be associated to any specific location.

4) Low recognition performance on report class could be the result from the characteristic of news report.

5.3.2.2 Probabilistic approach: highest probability feature

This experiment employed a more sophisticated approach by taking the statistical information generated from various features into consideration. The features used here are the 11 text-based features that are introduced in section 5.3.1, which are;

- Location names that appear in or modify subject
- Location names that appear in or modify subject's coreference

- Location names that appear in or modify object
- Location names that appear in or modify object's coreference
- Location names that appear in the phrase that directly modifies verb
- Location names that appear in the phrase that directly modifies verb's coreference
- Inference or co-reference of locations that appear in verb modifier
- News agency location
- Textually Closest location names
- Location in headline
- Previous verb's location.

Moreover, another feature is also included, which is the location names that appear in the same sentence as in-focus verb, as a feature for spatial attribute recognition.

In the experiment, the corpus was separated into training and testing corpus. In the training corpus, the locations generated from each feature, if available, are extracted. Then, the score of each feature type is calculated according to the following equation.

$$Score(f_i) = \frac{fc_i}{fa_i} * \left(\frac{1}{\log(fa_i - fc_i) + 1} \right)$$

where, fc_i is the number of times feature f_i gives the correct answer, and fa_i is the number of times that feature f_i gives the answer.

Intuitively, this scoring function is a weighted conditional probability; the feature receives a high score if it gives a correct spatial attribute to an event. In this scoring, I would like to give greater weight to the feature that has smaller difference between f_a and f_c, so small penalty in the form of log of the difference between f_a and f_c was used as a weight to the conditional probability.

For each event in the test set, the locations of each feature, if available, are extracted. Among the features that are fired (i.e., provide the geographical locations), the location generated from the feature that has the highest score is selected as the location of the event. The algorithm for calculating feature score is illustrated in figure 5-2.

Suppose: $EVENT = \{ ev \mid ev \in$ events in DOC $\}$
$loc_{ans} = \{l \mid l \in$ spatial attribute of $ev\}$
$FS = \{fs_j \mid fs_j \in$ predefined textual feature source$\}$
$DOC = \{d_i \mid d_i \in$ Training corpus$\}$

CalculateScore
Input: *EVENT*, *FS*, *DOC*
Output: Score of each feature in *FS*

For each *ev* in *EVENT*
For each feature source fs_j in *FS*
$loc_j = \{l' \mid l' \in$ Extracted_Location(fs_j, *ev*)$\}$
if loc_j is not null
$fa_j = fa_j + 1$
if member(loc_j) is exactly the same as member(loc_{ans})
$fc_j = fc_j + 1$
else if member(loc_j) is partially the same as member(loc_{ans})
$fc_j = fc_j + 0.25$

For each fs_j in *FS*
fs_j_score = Score(fs_j)

Figure 5-2: Algorithm for calculating feature score

Experimentation results

The results of the attribute annotation according to this approach are shown in the Table 5-5. As in the previous experiment, the annotation scores were also separated according to the event class.

Table 5-5: Experimentation results for recognizing spatial attribute of the events based on the probabilistic approach

Features	Event type	Precision (%)	Recall (%)	F (%)
11 textual features, without feature from previous verb's spatial attribute	Normal	76.4	74.8	75.6
	Report	87.0	85.7	86.3
	Information	63.5	65.6	64.5
	Hypothetical	65.9	65.9	65.9
	Overall	75.7	76.6	76.2
11 textual features, with feature from previous verb's spatial attribute	Normal	79.5	70.8	74.9
	Report	76.0	66.8	71.1
	Information	37.8	34.3	36.0
	Hypothetical	54.5	58.5	56.5
	Overall	64.3	71.8	67.9

From the results, we can see that when incorporating feature from previous verb's location that was predicted by the model, the results is apparently lower than without employing this feature. This could be the consequence from error propagation in recognizing event's location. Incorrectly recognized location of previous verb will have an impact in recognizing location of the current verb if the highest-probability feature is the location of the previous verb.

Separately analysis based on each event type, these results showed that the precision and recall of normal and reported events significantly improve compare to the results from heuristic approach. However, the results on information and hypothetical events are still quite low. This could indicate that in order to correctly recognize locations of information and hypothetical events, a deeper knowledge may be necessary. Incorporation of other information, such as event class or type of subject might help improving the performance. For example, sentence that discussed about the description of the disease usually consider as world knowledge, i.e. consider "world" as spatial attribute of the event represented by the verb in that sentence.

5.3.2.3 Statistical machine learning approach

As discussed earlier, there is usually a certain set of textual features that human readers used as an information source for signaling location of the events. In this experiment, I asked annotator A to annotate each event expression in the corpus. However, in stead of annotating the spatial attribute, the annotator was asked to select from the list of 11 textual feature (See section 5.3.1) that is the most likely to signal the spatial attribute of each event.

To develop an automatic system, I employ various statistical machine learning technique, which are Conditional Random Fields (CRF) [130], Decision Tree [131], and Support Vector Machine (SVM) [132], to learn to select the most likely feature to give the correct spatial attribute given a certain textual context. I employ CRF++[5] as a tool for CRF learning, libsvm[6] for SVM learning [133], and C4.5[7] for Decision tree learning [134]. The location extracted from the feature that the model selected will be considered

[5] http://crfpp.sourceforge.net/
[6] http://www.csie.ntu.edu.tw/~cjlin/libsvm
[7] http://www.rulequest.com/Personal/

as the event's spatial attribute predicted by the model. Parameter settings for each machine learning model are as follows:

For CRF++, *f* and *c* parameters were set as 3 and 4, respectively.

For SVM, linear kernel was used.

For C4.5, the default setting was used.

As mentioned in the agreement analysis of spatial attribute annotation, there are some certain sets of event expressions that usually cause disagreement between human annotator in selecting geographical locations as event's spatial attribute. Therefore, it is possible that other spatial-irrelevant information, such as class of event, may have an impact on spatial attribute annotation. Base on this hypothesis, not only the 11 textual feature sources, other information, such as event class, or type of subject are also incorporated into the model in order to evaluate the relevance of these information to the spatial attribute annotation task.

In the training task, I encode the value of each feature according to what are shown in Table 5-6

Table 5-6: Feature encoding method for the learning task

Feature	Value
11 textual features (according to section 5.3.1)	If the geographical locations can be extracted from the feature f_i then the value of f_i is encoded as "Y"; Otherwise the value of f_i is encoded as "N"
Event class	"Normal", "Report", "Information", "Hypothetical"
Subject type	Category of subject, which are: "Disease_Germ" (i.e. disease or pathogen) "Symptom" (i.e. symptom of any disease) "Officer" (i.e. person who has the property of being an official, such as government or medical officer) "Civilian" (i.e. person who does not have the property of being an official) "Gov. Organization" (i.e. government organization) "WHO" (i.e. World Health Organization) "WHO-related" (i.e. WHO agency or UN agency) "Organization" (i.e. other kinds of organization) "Location" "Other" (none of the above class)

Experimentation results

In this experiment, the learning models were trained to imitate how human utilize the information sources for recognizing spatial attribute of the event. The place where the event occurred was select according to the location that provided by the model-selected information source. The results of the attribute annotation according to this approach are shown in the Table 5-7. The scores for overall performance were based on micro-averaging. Note that the event class used as a feature in the model is the result from an automatic event classification system.

Table 5-7: Experimentation results for recognizing spatial attribute of the events based on the statistical machine learning approach

Feature	Class	CRF	SVM	C4.5
11 spatial-related textual features	Normal	84.8 (85.8, 83.8)	84.6 (85.8, 83.5)	78.75 (79.6, 77.9)
	Report	82.65 (81.85, 83.5)	83.7 (82.7, 84.6)	85.7 (85.2,86.2)
	Information	62.6 (63.0, 62.2)	54.8 (55.2, 54.5)	44.3 (45.0, 43.6)
	Hypothetical	88.1 (92.5, 84.1)	80.95 (85.0, 77.3)	64.4 (65.1, 63.6)
	Over all	81.3 (81.9, 80.7)	80.0 (80.7, 79.4)	74.9 (75.5, 74.4)
11 spatial-related textual features + event class	Normal	87.2 (88.3, 86.1)	86.6 (87.7, 85.5)	82.8 (84.05, 81.6)
	Report	86.1 (85.6, 86.6)	83.7 (82.7, 84.6)	88.4 (87.9, 88.9)
	Information	68.2 (68.4, 67.9)	65.8 (66.2, 65.4)	55.3 (56.8, 53.8)
	Hypothetical	76.2 (80.0, 72.7)	78.6 (82.5, 75.0)	78.2 (79.1, 77.3)
	All	83.8 (84.5, 83.1)	82.6 (83.3, 82.0)	80.1 (81.0, 79.2)
11 spatial-related textual features + subject type	Normal	85.7 (87.4, 84.1)	84.8 (86.4, 83.2)	79.5 (80.8, 78.3)
	Report	85.8 (84.9, 86.6)	83.7 (82.7, 84.6)	87.2 (86.3, 88.0)
	Information	75.5 (76.0, 75.0)	58.7 (59.1, 58.3)	61.3 (61.1, 61.5)
	Hypothetical	80.95 (85.0, 77.3)	83.3 (87.5, 79.5)	61.2 (61.9, 60.5)
	Over all	84.1 (85.1, 83.1)	80.75 (81.6, 79.9)	78.0 (78.5, 77.4)

11 spatial-related textual features + subject type + event class	Normal	86.7 (88.1, 85.4)	86.1 (87.8, 84.4)	80.2 (80.9, 79.5)
	Report	87.1 (86.4, 87.8)	84.0 (83.1, 85.0)	88.0 (87.45, 88.5)
	Information	80.4 (80.6, 80.1)	68.4 (68.8, 67.9)	66.7 (68.7, 64.7)
	Hypothetical	76.2 (80.0, 72.7)	81.0 (85.0, 77.3)	64.4 (65.1, 63.6)
	Over all	85.5 (86.3, 84.7)	82.9 (83.8, 82.0)	79.5 (80.1, 78.8)

Comparing between the three machine learning techniques, CRF yielded the best performance (F=87.2%). From the results, the CRF model, incorporating location-related features, subject type, and event class gained the highest performance.

While the CRF and SVM models with all features performed the best, incorporating only 11 location-related features and event class yielded a better result for Decision tree (F=80.1%). Although the subject type feature obviously improved the performance for CRF (from F=81.3% to 84.1%) and Decision tree model (from F=74.9% to 78.0%), it has slight positive impact to the SVM model (from F=80.0% to 80.8%). Nevertheless, the result from the models trained by the three machine learning techniques indicates that the event class and subject type features are useful for recognizing the spatial attribute of events in the information class.

In order to study the necessity and the contribution of each spatial-related textual feature, another set of experiment was conducted. In this experiment, each spatial-related textual feature was removed from the model in excluding-one-feature-per-training manner. Remark that the event class and subject type features were included in all models. The results were shown in detail in the Table 5-8 and were summarized in the graphical form in the Figure 5-3.

Table 5-8: Experimentation results of spatial attribute recognition from the models trained with different combinations of spatial-related textual features

Feature	Class	CRF	SVM	C4.5
Model I Without spatial information from subject	Normal	84.3 (85.7, 82.9)	82.9 (84.6, 81.2)	72.8 (77.9, 68.4)
	Report	81.4 (81.7, 81.1)	78.9 (78.3, 79.5)	78.7 (82.05, 75.6)
	Information	78.1 (78.6, 77.6)	67.3 (68.0, 66.7)	69.1 (72.5, 66.0)
	Hypothetical	71.4 (75.0, 68.2)	80.95 (85.0, 77.3)	77.6 (80.5, 75.0)
	Over all	82.2 (83.3, 81.1)	79.65 (80.7, 78.6)	*73.9* (78.2, 70.0)
Model II Without spatial information from subject coreference	Normal	85.1 (87.6, 82.7)	84.5 (86.6, 82.6)	76.3 (80.7, 72.3)
	Report	84.4 (84.9, 83.9)	80.4 (80.1, 80.7)	80.7 (82.4, 79.1)
	Information	76.8 (77.3, 76.3)	67.1 (67.5, 66.7)	69.6 (72.7, 66.7)
	Hypothetical	83.3 (87.5, 79.5)	88.1 (92.5, 84.1)	72.3 (76.9, 68.2)
	All	83.7 (85.4, 82.0)	81.2 (82.5, 80.0)	76.2 (79.8, 72.9)
Model III Without spatial information from object (and indirect object)	Normal	84.9 (86.2, 83.7)	84.0 (82.3, 85.85)	78.2 (81.9, 74.7)
	Report	86.1 (85.6, 86.6)	83.4 (82.6, 84.25)	87.0 (87.0, 87.0)
	Information	76.9 (76.9, 76.9)	65.6 (65.4, 65.8)	65.8 (68.3, 63.5)
	Hypothetical	74.7 (79.5, 70.45)	81.9 (87.2, 77.3)	76.7 (78.6, 75.0)
	Over all	83.7 (84.5, 82.9)	81.1 (82.2, 80.2)	78.5 (81.1, 76.0)
Model IV Without spatial information object (and indirect object) coreference	Normal	87.0 (88.3, 85.7)	86.0 (87.5, 84.4)	76.7 (78.1, 75.3)
	Report	88.3 (87.55, 89.0)	83.7 (82.7, 84.6)	87.35 (87.7, 87.0)
	Information	79.6 (80.1, 79.0)	68.4 (67.9, 68.8)	62.9 (65.1, 60.9)
	Hypothetical	77.5 (81.6, 73.8)	83.3 (87.5, 79.5)	66.7 (67.4, 65.9)
	Over all	**85.9** (86.75, 85.1)	82.2 (83.7, 82.0)	76.9 (78.2, 75.6)
Model V Without Nearest location feature (both nearest country-level locations and nearest city-level locations)	Normal	79.7 (81.3, 78.3)	81.65 (82.75, 80.6)	78.4 (75.3, 81.8)
	Report	87.3 (86.8, 87.8)	86.5 (85.7, 87.4)	87.4 (89.0, 85.8)
	Information	69.05 (70.2, 67.9)	64.7 (66.0, 63.5)	64.2 (67.1, 61.5)
	Hypothetical	57.5 (58.1, 56.8)	54.8 (57.5, 52.3)	47.1 (48.8, 45.45)
	Over all	79.1 (80.1, 78.15)	79.4 (80.2, 78.6)	77.3 (80.1, 74.6)

Model VI Without spatial information from headline	Normal	86.7 (87.95, 85.5)	86.6 (88.2, 85.05)	80.8 (79.1, 77.3)
	Report	87.1 (86.4, 87.8)	84.0 (83.1, 85.0)	88.6 (88.3, 89.0)
	Information	79.35 (79.9, 78.8)	68.4 (68.8, 67.9)	69.5 (71.9, 67.3)
	Hypothetical	78.6 (82.5, 75.0)	83.3 (87.5, 79.5)	78.2 (79.1, 77.3)
	Over all	85.5 (86.2, 84.7)	83.3 (84.2, 82.4)	**80.95** (82.3, 79.6)
Model VII Without spatial information about news agency	Normal	86.25 (85.05, 87.5)	86.0 (87.5, 84.4)	79.9 (80.9, 78.9)
	Report	87.1 (86.4, 87.8)	83.7 (82.7, 84.6)	87.6 (87.45, 87.8)
	Information	78.7 (79.2, 78.2)	68.4 (68.8, 67.9)	66.7 (68.7, 64.7)
	Hypothetical	78.6 (82.5, 75.0)	83.3 (87.5, 79.5)	64.4 (65.1, 63.6)
	Over all	85.1 (85.9, 84.3)	82.8 (83.7, 82.0)	79.2 (80.1, 78.3)
Model VII Without spatial information of previous event expression	Normal	79.1 (80.6, 77.65)	76.6 (78.5, 74.7)	80.1 (82.1, 78.3)
	Report	87.7 (87.2, 88.2)	86.5 (85.7, 87.4)	86.4 (87.8, 85.0)
	Information	63.2 (64.2, 62.2)	59.1 (59.9, 58.3)	64.0 (66.0, 62.2)
	Hypothetical	78.2 (79.1, 77.3)	71.3 (72.1, 70.45)	75.9 (76.7, 75.0)
	Over all	*78.8* (79.8, 77.9)	*76.2* (77.3, 75.2)	79.1 (80.9, 77.4)
Model IX Without spatial information of verb	Normal	83.5 (85.0, 82.0)	79.7 (81.95, 77.7)	76.4 (79.7, 73.3)
	Report	84.5 (84.3, 84.6)	81.2 (80.9, 81.5)	82.3 (85.2, 79.5)
	Information	77.0 (77.8, 76.3)	64.9 (65.8, 64.1)	63.3 (66.0, 60.9)
	Hypothetical	78.6 (82.5, 75.0)	84.7 (87.8, 81.8)	70.6 (73.2, 68.2)
	Over all	82.6 (83.7, 81.5)	78.2 (79.6, 76.8)	75.7 (78.8, 72.8)
Model X Without spatial information of event coreference	Normal	86.4 (87.8, 85.05)	85.4 (86.7, 84.1)	76.5 (78.9, 74.3)
	Report	86.7 (86.0, 87.4)	87.9 (86.9, 89.0)	82.9 (83.9, 81.9)
	Information	78.1 (78.6, 77.6)	69.0 (69.5, 68.6)	66.4 (69.0, 64.1)
	Hypothetical	80.95 (85.0, 77.3)	85.7 (90.0, 81.8)	64.4 (65.1, 63.6)
	Over all	85.1 (85.9, 84.2)	**83.7** (84.4, 82.95)	76.1 (78.1, 74.2)
Model XI Without spatial information from co-	Normal	86.7 (87.8, 85.7)	85.6 (87.1, 84.1)	78.8 (80.4, 77.2)
	Report	86.9 (86.4, 87.4)	83.4 (82.6, 84.25)	88.5 (88.9, 88.2)

reference of locations that appear in verb modifier	Information	77.4 (77.9, 76.9)	67.7 (68.2, 67.3)	70.6 (72.8, 68.6)
	Hypothetical	78.6 (82.5, 75.0)	83.3 (87.5, 79.5)	79.1 (80.95, 77.3)
	Over all	85.1 (85.9, 84.4)	82.45 (83.3, 81.6)	79.9 (81.4, 78.5)

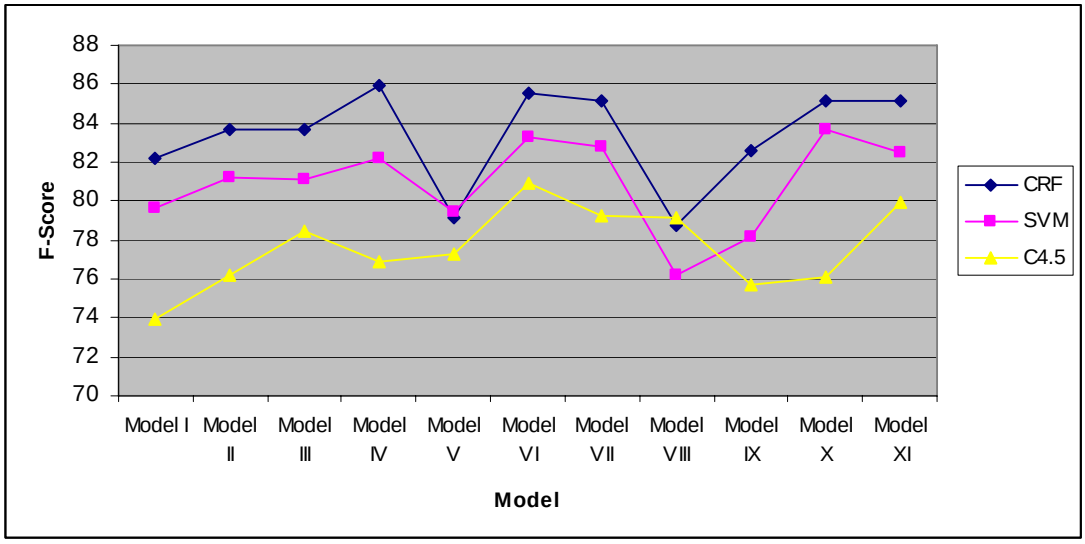

Figure 5-3: Spatial attribute recognition results from the models trained with different combinations of spatial-related textual features

According to the results, compared to the model trained with all the features (cf. Table 2), removing the headline feature improved recognition performance for the SVM (from 82.8 to 83.3 F-score) and Decision tree models (from 80.1 to 80.95 F-score), while yielded the same level of performance for the CRF model. Excluding the feature from the event coreference also improved the performance of the SVM model, but degraded the performance of the Decision tree model. For the CRF model, removing the object/indirect object coreference feature raised the recognition performance to 85.9, compared to the 85.5 F-score in the model that was trained with all the features. In CRF and SVM models, the recognition performance was most reduced when excluding the previous event feature (F=78.8% for CRF; F=76.2% for SVM). On the other hand, the performance of the Decision tree model was the lowest when excluding the location-related subject feature (F=73.9%).

Spatial attribute recognition performance was also analyzed for each event class. For events in the normal class, by removing the nearest locations and the previous event

feature cause a significant drop in performance for the CRF model (F=79.7% for CRF and 79.1% for SVM).

For events in the normal class, by removing the nearest locations and the previous event feature cause a significant drop in performance for the CRF model (F=79.7% for CRF and 79.1% for SVM). Among the SVM models, removing the previous event features has the most impact on the spatial attribute recognition performance (F=76.6%). Although the trend of the performance is quite similar between SVM and CRF models, the results from the decision tree model are obviously different. For the decision tree-trained model, removing the location-related subject feature has the most impact on the recognition performance (F=72.8%).

For the spatial attribute recognition of reporting events, the results show that the performance was reduced the most when the spatial information of the subject and subject co-reference was removed from the model (F=81.4% for CRF; F=78.9% for SVM, and F=78.7% for Decision tree). This situation is reflected in all the models trained by using the three techniques.

Next, the spatial attribute recognition performance on the information class was analyzed. For the CRF and SVM models, removing the previous event feature has the largest negative impact on recognition performance (F=63.2% for CRF; F=51.9% for SVM). While removing the object and indirect object coreference features has almost no impact on the CRF and SVM models in recognizing the spatial attribute of the information class, it surprisingly causes the most drop in performance for the decision tree model (F=62.9%).

For the spatial attribute recognition of hypothetical events, without taking the nearest locations into consideration, the recognition of the spatial attribute of the event in the hypothetical class was significantly reduced in all the models trained by using the three techniques.

Error analysis

A detailed data investigation was conducted in order to find the main source of errors in spatial attribute recognition. Based on the analysis, it was found out that the main causes of the errors can be grouped into 5 cases, which are:

Case1: Incorrect event class prediction

Most errors in this group occurred when an event in the normal class was classified by the model as an information class or vice versa. A major subcategory of the information class is a generic event or world knowledge, which, in this work, I regarded its spatial attribute as "world". When a normal event was miss-classified as information, it was often found that "world" was incorrectly recognized as the spatial attribute of the event.

Case2: Error propagation

Errors sometimes occurred when the most reliable source for spatial attribute annotation of the current event relies on the predicted locations of another event, such as a previous event location, event coreference's location or a subject coreference's location that was identified according to the model-predicted location of the event that it performed. For example, if the spatial attribute of the previous event, e_{i-1}, was incorrectly recognized, then the spatial attribute of the event e_i that relies on e_{i-1} will also be incorrectly recognized.

Case3: Confusion between selecting the Country-level and City-level nearest location

There usually is confusion, even for a human annotator, in deciding between selecting the location in the country-level or the lower level of administration as a spatial attribute of events. For example:

The villages, some 10 km north Jowhar lack Health Care centers and children have been dying from contagious deceases for the last years as the official ***confirmed***.

The above example is the excerpt from an outbreak report for villages around Somalia's middle Shabele region. While the automatic system selected Jowhar as the spatial attribute, according to the annotation guidelines, the preference is given to Shabelle region to be selected as the spatial attribute of the event "confirmed".

Base on the data investigation, this type of error occurred most often with the reporting events.

Case4: Spatial attribute recognition for information class

Information class can be subcategorized into generic knowledge and non-eventive clauses (for example, "The patient is a 6-year-old boy."). While the spatial attribute of generic knowledge is usually annotated as world or sometimes as a country-level

location, the spatial attribute of a non-eventive clause is generally a specific location. For example:

It *is* the first time something like this is happening at our school.

The non-eventive information represented by the phrase “is the first time …” should be anchored to the location of the school. Or,

Local Community Health Care in the area told AFP that lack of health care *is* the main cause of the amazing children death number.

The clause “lack of health care is the main cause ...” refers to the fact in a specific region, (which, in this news, is "Shabele"), so Shabele was intuitively selected by the human annotator as the location of this non-eventive clause.

It is quite difficult to find explicit textual signals that can guide a model to select the most appropriate feature source for spatial attribute annotation. However, it is often found that, in cases where the event is classified as information, when the subject of the verb referred to a non-specific, concept-level entity, such as "Bird flu", "Patients", "Children", the event represented by that verb is usually considered world knowledge, i.e. the spatial attribute is "World". Based on this observation, an analysis of a subject whether it refers to a concept level entity or not might help solve this error.

Case5: Event that causes the spatial movement of the object

When the event involves the movement of an object from one place to another, both the source and destination locations should be recognized as the spatial attribute of the event. For example:

There are fears the disease *could spread* into neighboring Uganda.

This news reported an outbreak in Sudan, so the hypothetical event “could spread” should be associated with a source location, e.g. Sudan, and the destination location, e.g. Uganda.

However, this type of event seldom occurred in the training corpus. So, it is difficult for the model to recognize this.

5.3.3 Discussion

From the results, the system developed based on statistical machine learning approach has the highest performance in event spatial attribute annotation. However, in the

experiment, I manually identified all subject and object of every verb, as well as the co-reference information in order to avoid the impact of the parsing and co-reference resolution performance to the spatial attribute annotation task. For the practical use, these information must also be automatically done by using available linguistic tool, such as parser [135], co-reference resolution [136], etc. Although the error in subject/object identification, as well as coreference resolution, could cause degradation in the annotation performance, with the current advance in natural language processing the results of spatial attribute annotation should still be reliable enough to be used in the Web-based health surveillance system.

5.4 Automatic approach for temporal attribute annotation

In this work, the task of temporal annotation was considered as the task to identify temporal relation between time and event that was explicitly mentioned in text. The temporal relation could be recognized from:

1) Direct relation between event and temporal expression

This is the general form of relation that the system must recognize. Generally, the temporal relations in this group can be detected through a certain set of linguistic signals that appear between event and temporal expression. These linguistic signals are, such as the presence of adverb or preposition “at”, “in”, “before”, “after”.

2) Inference from explicitly-mentioned temporal relation between events

Temporal relation can also be inferred when the temporal relation is explicitly mentioned between two events, trough the usage of temporal relation signal, such as “before”, “at the same time”, “after”, and the time of one of the events in the relation is known. For example:

> On Christmas day, a 24-year-old woman from Jakarta also **died** from the virus after **buying** a live chicken from a market.

In this situation, the system must be able to infer the temporal relation between the event “buying” and “Christmas day”.

3) Inference from event co-reference information

In news report, it is often that the same event is repeatedly mentioned many times, while the occurring time of the event is stated only once. For example:

> "While the 17 people are stable, they are not yet out of danger," Medical Service director-general Dr. Chatri Banchuen **said** yesterday.
> "Basically, their condition remains unsafe," he **added**.

Although there is no explicit temporal relation signal between event "added" and the time "yesterday", or between "said" and "added", but general readers can easily infer that the two events were consecutively occurred or co-referred to the same situation where the doctor gave an interview to the media. It is quite intuitive that "said" and "add" occurred in the same day, which is "yesterday". In this circumstance, the system should be able to infer temporal relation between the event, "added", and time, "yesterday".

In order to recognize temporal relations, either between event and event or event and temporal expression, various approaches have been proposed. Grammatical rule-based strategy was the straightforward method to capture temporal relation. Since these relations are usually expressed through linguistic signal and grammatical relation. Many works viewed the problem of recognizing temporal relation as a classification problem, where the task is to assign a label to temporal link. The machine learning techniques that were used are, such as Hidden Markov Support Vector Machine (HMM_SVM), Support Vector Machine (SVM), Naïve Bayes. In this thesis, a simple rule-based approach is employed.

5.4.1 Temporal attribute recognition approach

The temporal relations to be captured in this work are those that were explicitly mentioned in text, i.e. generally signaled by linguistic clues. Since the information sources for recognizing temporal relations were mainly based on grammatical and linguistic information, as well as a small number of training corpus, the rule-based approach was selected for performing this task.

In order to avoid the impact from word-ordering to the rule construction task, rules for recognizing temporal attributes were coded from grammatical dependency between each sentence elements. The input documents of the annotation system were: 1) named entity marked-up, 2) verb phrase marked-up, and 3) parsed by dependency parser[8] [137].

[8] Stanford parser, available at http://nlp.stanford.edu/software/lex-parser.shtml

Examples of grammatical rules employed for temporal attribute annotation are shown in the table below.

Table 5-9: Examples of rules for recognizing event's temporal attribute

Pre-assumption: *v* is the head of verb phrase *t* is a token element in the temporal expression, *t_exp* *t′* is a token element in the temporal expression *t_exp′*		
Rules	Stime	Etime
prep_from(*v*, *t*) prep_since (*v*, *t*)	Stime <- *t_exp* Stime_dir <- "AS_OF"	
prep_to(*v*, *t*) prep_until(*v*, *t*) prep_till(*v*, *t*) prep_by(*v*, *t*)		Etime <- *t_exp* Etime_dir <- "AS_OF"
prep_in(*v*, *t*) prep_at(*v*, *t*) prep_on(*v*, *t*)	Stime <- *t_exp* Stime_dir <- "AS_OF"	Etime <- *t_exp* Etime_dir <- "AS_OF"
prep_between(*v*, *t*) conj_and(*t*, *t′*)	Stime <- *t_exp* Stime_dir <- "AS_OF"	Etime <- *t_exp′* Etime_dir <- "AS_OF"
prep_before(*v*, *t*)	Stime <- *t_exp* Stime_dir <- "BEFORE"	
prep_after(*v*, *t*)	Stime <- *t_exp* Stime_dir <- "AFTER"	

As mentioned earlier, it is usually found in newspaper that various expressions that co-refer to the same event (or macro-event) is repeatedly mentioned many times, while the occurring time of the event is stated only once. This situation occurred quite often especially for reporting events. In order to improve the performance of the system in recognizing temporal information of such events, simple heuristic is used for identifying the linguistic expressions that refer to the same events. The event expressions that were expressed by; 1) the same verbs or verbs with the same sense, and 2) have the subjects that co-refer to the same entity; will be considered as the co-referring events.

Co-referring information between events will be used for assigning extracted temporal attribute of one event to the rest of the co-referring events.

There are also another set of grammatical rules for managing the case where the two events hold explicit temporal relation to each other. The rules in this set attempt to recognize the pattern in order to assign the absolute temporal attributes to the event according to the temporal attributes that were extracted for the other event and the temporal relation between the two events.

5.4.2 Temporal attribute recognition: Experimentation

The experimentation task consists of two steps, which are rule development task, and temporal attribute annotation task. In rule development process, rules were encoded from the corpus set2. These rules were tested on the news articles in corpus set1.

In the corpus set1, all 1086 events, only 215 events were manually annotated with absolute temporal information. The rest of 871 events were annotated according to verb tense, which were considered as an easy task for automatic system to annotate such temporal information. In order to reflect the system performance in annotating absolute temporal information, the annotation results on annotating event with absolute time are presented separately. The results of temporal attribute annotation are shown in table 5-10.

Table 5-10: Experimentation results of temporal attribute annotation

Rules		Precision	Recall	F-score
Without event co-referring information	All annotation	90.4%		
	Absolute time annotation	89.6%	55.8%	68.8%
With event co-referring information	All annotation	93.3%		
	Absolute time annotation	90.5%	70.4%	79.2%

According to error investigation, many errors occurred when the temporal information were stated as a duration, such as "during the previous three days", "the last two months". These expressions were not recognized as temporal entities based on BioCaster's named entity annotation guideline. Consequently, the rules can not detect temporal information of the events that linked to durative temporal expressions. In order to improve the performance of the system, the ability to handle these durative temporal expressions is necessary. One way to solve this problem is to adopt the temporal expression annotation from TIMEX3 specification [98]. The difficulty in capturing inter-sentence temporal relation by rules also caused the problem in temporal attribute annotation. For example:

The Ministry of Health has reported to WHO that 41 cases from 18 cities and provinces **have been detected** in Viet Nam since *mid-December 2004*. Of these cases, 16 **have died** and six **remain under treatment**.

In the above example, the first example discussed about the detection of 41 cases, while the second sentence gave more detail about the health condition of the same group of victims. To recognize this kind of temporal relation, apart from grammatical rules, other feature sources may be necessary.

The parsing error is also another source that causes the incorrect recognition of temporal attribute.

Chapter 6

Conclusion in Future Work

This thesis described a novel framework called spatiotemporal zoning for the purpose of enhancing the ability of Web-based health surveillance systems in recognizing spatial and temporal information of events in a finer granularity. The literature has pointed out that a fine understanding of temporal and spatial information about events, either directly or indirectly related to outbreak situation, in outbreak reports is necessary for situation analysis, decision making, or risk management. Manual annotation experiments showed that the proposed scheme is reproducible and can be learned by human annotators. However, to be of practical use in a real system, the scheme needs to be automated. To be of practical use in a real health surveillance system, various approaches and features have been investigated for automatically analyzing text content according to the spatiotemporal scheme. The results of the automatic zone annotation were quite promising. The best performance for content class annotation was 89.2% F-score. For spatial attribute annotation, the best performance from machine learning model was 85.9% F-score. The temporal attribute annotation, however, was affected by parsing errors and situations that could not be covered by grammatical rules. These problems can be overcome by increasing the number of rules and incorporating a more sophisticated approach for identifying inter-sentence temporal relations. Moreover, since the temporal aspect of this work was based on TimeML framework, the continuity studies on automatic temporal annotation in the evaluation task such as TempEval may provide good resources for making improvements to the temporal attribute annotation.

In the experiment, the named entity, co-reference information and the subjects and objects of all verbs were manually identified. In practice, such information can be automatically acquired by using linguistic tools, such as named entity annotation [104], parser [137], and co-reference resolution [136]. Since the errors from these tools could degrade the annotation accuracy, it is necessary to have the tools with acceptable performance. One possible way to improve the performance of the parsers is to train them on a corpus in the same domain that was used in the experiment. Moreover, to my

knowledge, there is no off-the-shelf tool for coreference resolution that can provide all the information needed in the proposed methodology for spatiotemporal zone annotation. For an effective implementation of the zone annotation system, a coreference resolution tool needs to be developed. Nevertheless, with continual advances in natural language processing, reliable tools may soon appear to make the spatial attribute annotation reliable enough for actual online health surveillance systems

In terms of domain dependency, although this thesis focused on the disease outbreak domain, its characteristic of analyzing real-world events reported in text means that its results might also be valid to other domains that involve tasks of identifying the places and times of real-world events.

As mentioned in chapter 3, the temporal and spatial granularities were designed to satisfy the needs of health surveillance systems. For the domains that need more detailed information, the spatial and temporal granularities can be adjusted. For example, the temporal granularity can be defined at the time-of-day level for domains that require finer temporal information. Moreover, the current scheme defines spatial information about an event with one zone attribute indicating "occur in" relation between events and location filled in that attribute. However, for domains that involve motion-related events, such as "spread", "send", "move", etc., the notion of spatial attributes should be extended to accommodate both starting and ending locations of an event. This can be done by, for example, introducing two spatial attributes to indicate the source and destination locations.

This scheme was designed based on the notion of an event, which is language independent. That means the scheme itself can be applied to documents in any language. On the other hand, the methodology of automatic zone annotation proposed in this thesis may need to be modified to work in different languages. The availability of linguistic tools in various languages is also a limitation in utilizing the automatic zone annotation methodology.

Future studies

- Explore other methods to enhance the performance of zone annotation

To improve a performance of automatic zone annotation, other approaches, as well as features, should be explored. The event class annotation had a lot of mistakes in

distinguishing between events in normal class and information class. Knowledge and features deeper than surface form and syntactic information is needed to make such a distinction. For spatial attribute annotation, external features, such as spatial relation between locations, should also be investigated. Also, the temporal attribute annotation results revealed cases that could not be handled only by employing grammatical rules. Accordingly, approaches that combine grammatical information and heuristic or statistical information should be explored.

- Apply spatiotemporal zoning to a Web-based health surveillance system

To prove its usefulness, spatiotemporal zoning will have to be applied to actual health services tasks. In the future, I plan to apply spatiotemporal zoning to a health surveillance system in order to enhance the capability of the system in detecting outbreak places and time in a better granularity.

- Exploiting redundancy in news reports: Multi-document analysis

It is common for news stories about the same event to be published by various news agencies over the course of several days. Such articles may have different levels of spatiotemporal information about an event. That is, some articles may provide only brief (or even neglect) information regarding the place and time of an event, while others may provide very detailed spatiotemporal information. The ability to recognize the same event reported in different news articles could help to ameliorate situations wherein the system cannot acquire detail spatiotemporal information about an event within the same document. To enable this capability, a means of identifying the same event reported in different documents and linking these documents together, i.e. cross-document zoning, should be developed.

About Author

Name	Hutchatai Chanlekha
Birth Date	April 5, 1980
Birth Place	Bangkok, Thailand
Nationality	Thai
Status	Single
Educations	2001-2004 Master of Computer Engineering, Department of Computer Engineering, Kasetsart University, Bangkok, Thailand 1997-2001 Bachelor of Computer Engineering (First Class Honors), Department of Computer Engineering, Kasetsart University, Bangkok, Thailand
Experiences	Jun 2001 - Sep 2006 Research assistant, Department of Computer Engineering, Kasetsart University, Bangkok, Thailand

Related Publication

Chanlekha H, Collier N, Kawazoe A: **A Step Towards Disease Outbreak Information Extraction: Automatic Entity Role Recognition for Named Entities**. In: *Proc of the seventh International Symposium on Natural Language Processing.* Chonburi, Thailand; 2007.

Chanlekha H, Kawazoe A, Collier N: **A framework for enhancing spatial and temporal granularity in report-based health surveillance systems**. *BMC Medical Informatics and Decision Making* 2010, 10(1).

Chanlekha H, Collier N: **Analysis of syntactic and semantic features for fine-grained event-spatial understanding in outbreak news reports**. In: *Proc of the 3rd International Symposium on Languages in Biology and Medicine (LBM).* Jeju Island, South Korea; 2009.

Chanlekha H, Collier N: **A methodology to enhance spatial understanding of disease outbreak events reported in news articles**. *International Journal of Medical Informatics* 2010, DOI: 10.1016/j.ijmedinf.2010.01.014.

Chanlekha H, Collier N: **Analysis of syntactic and semantic features for fine-grained event-spatial understanding in outbreak news reports**. *Journal of Biomedical Semantics*. (To be appeared)

Bibliography

1. World Health Organization: **International Health Regulations (2005)**, 2 edn: World Health Organization; 2008.

2. Lewis MD, Pavlin JA, Mansfield JL, O'Brien S, Boomsma LG, Elbert Y, Kelley PW: **Disease outbreak detection system using syndromic data in the greater Washington DC area**. *American Journal of Preventive Medicine* 2002, **23**(3):180-186.

3. Tsui F-C, Espino JU, Dato VM, Gesteland PH, Hutman J, Wagner MM: **Technical Description of RODS: A Real-time Public Health Surveillance System**. *Journal of American Medical Informatics Association* 2003, **10**(5):399-408.

4. Heymann DL, Rodier GR: **Hot spots in a wired world: WHO surveillance of emerging and re-emerging infectious diseases**. *The Lancet Infectious Diseases* 2001, **1**(5):345-353.

5. Brownstein JS, Freifeld CC: **HealthMap: the development of automated real-time internet surveillance for epidemic intelligence**. *Eurosurveillance* 2007, **12**(48).

6. Butler D: **Disease surveillance needs a revolution**. *Nature* 2006, **440**(7080):6-7.

7. Yangarber R, Steinberger R, Best C, Etter Pv, Fuart F, Horby D: **Combining Information Retrieval and Information Extraction for Medical Intelligence**. In: *Proceeding of Mining Massive Data Sets for Security, NATO Advanced Study Institute.* Gazzada, Italy; 2007.

8. Mawudeku A, Blench M: **Global Public Health Intelligence Network (GPHIN)**. In: *Proceeding of the 7th Conference of the Association for Machine Translation in the Americas.* Cambridge, Massachusetts, United States of America; 2006: 7-11.

9. Mawudeku A, Lemay R, Werker D, Andraghetti R, John RS: **The Global Public Health Intelligence Network**. In: *Infectious Disease Surveillance.* Edited by M'ikanatha NM, Lynfield R, Beneden CAV, Valk Hd: Infectious Disease Surveillance; 2007: 304-317.

10. Wilson JM: **Argus: A Global Detection and Tracking System for Biological Events**. *Advances in Disease Surveillance* 2007, **4**:21.

11. Tolentino H, Kamadjeu R, Fontelo P, Liu F, Matters M, Pollack M, Madoff L: **Scanning the Emerging Infectious Diseases Horizon-Visualizing ProMED Emails Using EpiSPIDER**. *Advances in Disease Surveillance* 2007, **2**(4):169.

12. Collier N, Doan S, Kawazoe A, Goodwin RM, Conway M, Tateno Y, Ngo Q-H, Dien D, Kawtrakul A, Takeuchi K *et al*: **BioCaster: detecting public health rumors with a Web-based text mining system**. *Bioinformatics* 2008, **24**:2940-2941.

13. Collier N, Kawazoe A, Doan S, Shitematsu M, Taniguchi K, Jin L, McCrae J, Chanlekha H, Dien D, Hung Q *et al*: **Detecting Web rumours with a multilingual ontology supported text classification system**. *Advances in Disease Surveillance* 2007, **4**:242.

14. Keller M, Blench M, Tolentino H, Freifeld CC, Mandl KD, Mawudeku A, Eysenbach G, Brownstein JS: **Use of Unstructured Event-Based Reports for Global Infectious Disease Surveillance**. *Emerging Infectious Disease* 2009, **15**(5).

15. Morse SS: **Global Infectious Disease Surveillance And Health Intelligence**. *Health Affairs* 2007, **26**(4):1069-1077.

16. Brownstein JS, Freifeld CC, Reis BY, Mandl KD: **HealthMap: Internet-based emerging infectious disease intelligence**. In: *Global Infectious Disease Surveillance and Detection: Assessing the Challenges--finding Solutions: Workshop Summary.* National Academies Press; 2007: 183-204.

17. Grishman R, Huttunen S, Yangarber R: **Information extraction for enhanced access to disease outbreak reports**. *Journal of Biomedical Informatics* 2002, **35**(4):236-246.

18. Grishman R, Huttunen S, Yangarber R: **Real-time Event Extraction for Infectious Disease Outbreaks**. In: *Proceedings of the second international conference on Human Language Technology Research.* San Diego, California; 2002: 366-369.

19. Yangarber R, Best C, Etter Pv, Fuart F, Horby D, Steinberger R: **Combining Information about Epidemic Threats from Multiple Sources**. In: *Proceeding of the Workshop on Multi-source Multilingual Infor-mation Extraction and Summarization (MMIES'2007), RANLP'2007.* Borovets, Bulgaria; 2007.

20. Mykhalovskiy E, Weir L: **The Global Public Health Intelligence Network and Early Warning Outbreak Detection: A Canadian Contribution to Global Public Health**. *Canadian Journal of Public Health* 2006, **97**(1):42-44.

21. Moens M, Steedman M: **Temporal ontology in natural language**. In: *Proceeding of the 25th annual meeting on Association for Computational Linguistics.* Stanford, California: Association for Computational Linguistics; 1987.

22. Bestgen Y, Vonk W: **The Role of Temporal Segmentation Markers in Discourse Processing**. *Discourse Processes* 1995, **19**:385-406.

23. Blackburn P: **Tense, Temporal Reference, and Tense Logic** *Journal of Semantics* 1994, **11(1-2)**:83-101.

24. Dowty DR: **Tenses, time adverbs, and compositional semantic theory**. *Linguistics and Philosophy* 1982, **5**(Volume 5, Number 1 / March, 1982):23-55.

25. Dowty DR: **The effects of aspectual class on the temporal structure of discourse: semantics or pragmatics?** *Linguistics and Philosophy* 1986, **9**(Volume 9, Number 1 / February, 1986):37-61.

26. Hinrichs E: **Temporal anaphora in discourses of English**. *Linguistics and Philosophy* 1986, **9**(Volume 9, Number 1 / February, 1986):63-82.

27. Moens M, Steedman M: **Temporal ontology and temporal reference**. *Computational Linguistics: Special issue on tense and aspect* 1988, **14**(2):15-28.

28. Nakhimovsky A: **Temporal reasoning in natural language understanding: the temporal structure of the narrative**. In: *Proceeding of the Third Conference on European chapter of the Association for Computational Linguistics.* Copenhagen, Denmark 1987: 262-269.

29. Nelken R, Francez N: **Splitting the reference time:temporal anaphora and quantification**. In: *Proceedings of the EACL '95 - The seventh meeting of the European Chapter of the Association for Computational Linguistics.* 1995.

30. Partee BH: **Nominal and Temporal Anaphora**. *Linguistics and Philosophy* 1984, **7**(3):243-286.

31. Smith CS: **Temporal Structures in Discourse**. In: *Time, Tense and Quantifiers.* Edited by Rohrer C. Tubingen: Niemeyer; 1980: 355-374.

32. Pustejovsky J, Castaño JM, Ingria R, Sauri R, Gaizauskas RJ, Setzer A, Katz G, Radev DR: **TimeML: Robust Specification of Event and Temporal Expressions in Text**. In: *Proceeding of the Fifth International Workshop on Computational Semantics (IWCS-5).* 2003: 28-34.

33. Verhagen M, Gaizauskas R, Schilder F, Hepple M, Katz G, Pustejovsky J: **SemEval-2007 Task 15: TempEval Temporal Relation Identification**. In: *Proceedings of the 4th International Workshop on Semantic Evaluations.* Prague, Czech Republic: Association for Computational Linguistics; 2007.

34. Saurí R, Knippen R, Verhagen M, Pustejovsky J: **Evita: a robust event recognizer for QA systems**. In: *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing.* Vancouver, Canada: Association for Computational Linguistics; 2005: 700-707.

35. Mani I, Wellner B, Verhagen M, Pustejovsky J: **Three Approaches to Learning TLINKs in TimeML**. In. Waltham, USA: Computer Science Department, Brandeis University.; 2007.

36. Mani I, Verhagen M, Wellner B, Lee CM, Pustejovsky J: **Machine learning of temporal relations**. In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics.* Sydney, Australia: Association for Computational Linguistics; 2006: 753-760.

37. Tannier CHX: **XRCE-T: XIP temporal module for TempEval campaign**. In: *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007).* Prague, Czech Republic: Association for Computational Linguistics; 2007.

38. Puscasu G: **WVALI: Temporal Relation Identification by Syntactico-Semantic Analysis**. In: *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007).* Prague, Czech Republic: Association for Computational Linguistics; 2007.

39. Min C, Fowler MSA: **LCC-TE: A Hybrid Approach to Temporal Relation Identification in News Text**. In: *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007).* Prague, Czech Republic: Association for Computational Linguistics; 2007.

40. Bethard S, Martin JH: **CU-TMP: Temporal Relation Classification Using Syntactic and Semantic Features**. In: *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007).* Prague, Czech Republic: Association for Computational Linguistics; 2007.

41. Cheng Y, Asahara M, Matsumoto Y: **NAIST.Japan: Temporal Relation Identification Using Dependency Parsed Tree**. In: *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007).* Prague, Czech Republic: Association for Computational Linguistics; 2007.

42. Hepple M, Setzer A, Gaizauskas R: **USFD: Preliminary Exploration of Features and Classifiers for the TempEval-2007 Tasks**. In: *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007).* Prague, Czech Republic: Association for Computational Linguistics; 2007.

43. Collins M, Singer Y: **Unsupervised models for named entity classification**. In: *Proceeding of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.* 1999: 189-196.

44. Cucchiarelli A, Velardi P: **Unsupervised Named Entity Recognition Using Syntactic and Semantic Contextual Evidence**. *Computational Linguistics* 2001, **27**(1):123-131.

45. Borthwick A, Sterling J, Agichtein E, Grishman R: **NYU: Description of the MENE Named Entity System as Used in MUC-7**. In: *Proceeding of the 7th Message Understanding Conference.* Fairfax, Virginia; 1998.

46. Mikheev A, Moens M, Grover C: **Named Entity recognition without gazetteers**. In: *Proceeding of the ninth conference on European chapter of the Association for Computational Linguistics.* Bergen, Norway: Association for Computational Linguistics; 1999.

47. Niu C, Li W, Ding J, Srihari RK: **A Bootstrapping Approach to Named Entity Classification Using Successive Learners**. In: *Proceeding of the 41st Annual Meeting on Association for Computational Linguistics* vol. 1. Sapporo, Japan: Association for Computational Linguistics; 2003: 335-342.

48. Niu C, Li W, Srihari R, Crist L: **Bootstrapping a Hidden Markov Model for Relationship Extraction Using Multi-level Contexts**. In: *Proceeding of the Pacific Association for Computational Linguistics 2003 (PACLING03).* Halifax, Nova Scotia, Canada; 2003.

49. Baluja S, Mittal VO, Sukthankar R: **Applying machine learning for high performance named-entity extraction**. In: *Proceeding of Pacific Association for Computational Linguistics.* Waterloo, CA: Blackwell Publishers; 1999.

50. Appelt DE, Hobbs JR, Bear J, Israel D, Tyson M: **FASTUS: A Finite-state Processor for Information Extraction from Real-world Text**. In: *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI-93).* Chambery, France; 1993.

51. Weischedel R, Ayuso D, Bikel D, Bobrow R, Boisen S, Burstein M, Ferguson W, Fox H, Hyde C, Ingria R: **BBN: description of the PLUM system as used for MUC-6**. In: *Proceeding of the Sixth Message Understanding Conference (MUC-6).* Columbia, MD: Morgan-Kaufmann Publishers; 1995.

52. Stevenson M, Gaizauskas R: **Using corpus-derived name lists for named entity recognition**. In: *Proceedings of the sixth conference on Applied natural language processing.* Seattle, Washington: Association for Computational Linguistics; 2000.

53. Iwańska L, Croll M, Yoon T, Adams M: **Wayne State University: description of the UNO natural language processing system as used for MUC-6**. In: *Proceedings of the 6th Message Understanding Conference (MUC-6).* Columbia, Maryland: Association for Computational Linguistics; 1995.

54. Gaizauskas R, Humphreys K, Cunningham H, Wilks Y: **University of Sheffield: description of the LaSIE system as used for MUC-6**. In: *Proceedings of the 6th Message Understanding Conference (MUC-6).* Columbia, Maryland: Association for Computational Linguistics; 1995.

55. Wacholder N, Ravin Y, Choi M: **Disambiguation of proper names in text**. In: *Proceedings of the fifth conference on Applied natural language processing.* Washington D.C.: Association for Computational Linguistics; 1997.

56. Bikel DM, Miller S, Schwartz R, Weischedel R: **Nymble: a high-performance learning name-finder**. In: *Proceedings of the fifth conference on Applied Natural Language Processing* Washington, D.C.: Association for Computational Linguistics; 1997.

57. Borthwick A, Sterling J, Agichtein E, Grishman R: **Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition**. In: *Proceeding of the 6th Workshop on Very Large Corpora.* Montreal, Canada; 1998.

58. Borthwick A, Sterling J, Agichtein E, Grishman R: **NYU: Description of the MENE Named Entity System as Used in MUC-7**. In: *Proceeding of the 7tn Message Understanding Conference (MUC-7).* Fairfax, Virginia; 1998.

59. Borthwick A: **A Maximum Entropy Approach to Named Entity Recognition**. *Ph.D. thesis.* New York University; 1999.

60. Sekine S, Grishman R, Shinnou H: **A decision tree method for finding and classifying names in Japanese texts**. In: *Proceeding of the 6th Workshop on Very Large Corpora.* 1998.

61. Buchholz S, Bosch Avd: **Integrating seed names and n-grams for a named entity list and classifier**. In: *Proceeding of the Second International Conference on Language Resources and Evaluation (LREC 2000).* Greece; 2000.

62. Isozaki H, Kazawa H: **Efficient Support Vector Classifiers for Named Entity Recognition**. In: *Proceedings of the 19th international conference on Computational Linguistics.* Taipei, Taiwan: Association for Computational Linguistics; 2002.

63. Takeuchi K, Collier N: **Use of Support Vector Machines in Extended Named Entity Recognition**. In: *Proceedings of the 6th conference on Natural Language Learning (CoNLL-2002).* Taipei, Taiwan: Association for Computational Linguistics; 2002.

64. Collins M, Singer Y: **Unsupervised Models For Named Entity Classification**. In: *Proceeding of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.* Univ. of Maryland, MD; 1999.

65. Cucerzan S, Yarowsky D: **Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence**. In: *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.* Univ. of Maryland, MD; 1999.

66. Mikheev A, Grover C, Moens M: **Description of the LTG System Used for MUC-7**. In: *Proceeding of the 7th Message Understanding Conference (MUC-7).* Fairfax, Virginia; 1998.

67. Peng Y, He D, Mao M: **Geographic Named Entity Disambiguation with Automatic Profile Generation**. In: *Proceeding of IEEE/WIC/ACM International Conference on Web Intelligence (WI'06).* IEEE Computer Society 2006: 522-525.

68. Leidner JL, Sinclair G, Webber B: **Grounding spatial named entities for information extraction and question answering**. In: *Proceeding of HLT-NAACL 2003 workshop on Analysis of geographic references.* vol. 1: Association for Computational Linguistics; 2003: 31-38.

69. Bilhaut F, Charnois T, Enjalbert P, Mathet Y: **Geographic reference analysis for geographic document querying**. In: *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references.* vol. 1: Association for Computational Linguistics; 2003: 55-62.

70. Li H, Srihari RK, Niu C, Li W: **Location Normalization for Information Extraction**. In: *Proceedings of the 19th international conference on Computational linguistics.* vol. 1. Taipei, Taiwan; 2002: 1-7.

71. Volz R, Kleb J, Mueller. W: **Towards ontology-based disambiguation of geographical identifiers**. In: *Proceeding of the 16th International World Wide Web Conference (WWW2007).* Banff, Alberta, Canada; 2007.

72. Overell S, Rüger S: **Using co-occurrence models for placename disambiguation**. *International Journal of Geographical Information Science* 2008, **22**(3):265-287.

73. Chinchor N: **MUC-7 Information Extraction Task Definition (Version 5.1)**. In.; 1998.

74. Freitag D, McCallum A: **Information extraction with HMMs and shrinkage**. In: *Proceeding of the AAAI-99 Workshop on Machine Learning for Information Extraction.* Orlando, FL; 1999: 31-36.

75. Sun A, Naing M-M, Lim E-P, Lam W: **Using Support Vector Machines for Terrorism Information Extraction**. In: *Intelligence and Security Informatics.* vol. 2665/2003: Springer Berlin / Heidelberg; 2003.

76. Soderland S, Fisher D, Aseltine J, Lehnert W: **CRYSTAL: Inducing a Conceptual Dictionary**. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI).* Montreal, Canada; 1995.

77. Soderland S: **Learning Information Extraction Rules for Semi-Structured and Free Text**. *Machine Learning* 1999, **34**(1-3):233-272.

78. Freitag D: **Toward general-purpose learning for information extraction**. In: *Proceedings of the 17th international conference on Computational Linguistics.* Montreal, Canada: Association for Computational Linguistics; 1998.

79. Català N, Castell N, Martin M: **A Portable Methodology for Acquiring Information Extraction Patterns without Annotated Corpora**. *Natural Language Engineering* 2003, **9**(2):151-179.

80. Basili R, Pazienza MT, Vindigni M: **Corpus-driven learning of Event Recognition Rules**. In: *Proceedings of the ECAI Workshop on Machine Learning for Information Extraction.* 2000.

81. Yangarber R: **Counter-training in discovery of semantic patterns**. In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics.* Sapporo, Japan: Association for Computational Linguistics; 2003.

82. Riloff E: **Automatically Generating Extraction Patterns from Untagged Text**. In: *Proceeding of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)* vol. Volume 2. Portland, Oregon, United States: AAAI Press/MIT Press; 1996: 1044-1049.

83. Poibeau T, Dutoit D: **Generating extraction patterns from a large semantic network and an untagged corpus**. In: *Proceeding of the International Conference On Computational Linguistics COLING-02 on SEMANET: building and using semantic networks.* vol. Volume 11. Taipei, Taiwan: Association for Computational Linguistics; 2002.

84. Phillips W, Riloff E: **Exploiting Role-Identifying Nouns and Expressions for Information Extraction**. In: *Proceeding of the Conference on Recent Advances in Natural Language Processing (RANLP-2007)* Borovets, Bulgaria; 2007.

85. Riloff E: **Automatically constructing a dictionary for information extraction tasks**. In: *Proceeding of the Eleventh National Conference on Artificial Intelligence.* Washington, DC; 1993: 811-816.

86. Riloff E, Shepherd J: **A Corpus-Based Approach for Building Semantic Lexicons**. In: *Proceeding of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-97).* Providence, Rhode Island: Association for Computational Linguistics; 1997: 117-124.

87. Thelen M, Riloff E: **A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts**. In: *Proceeding of the second Conference on Empirical methods in natural language processing (EMNLP-02).* vol. 10. Philadelphia, PA, United States: Association for Computational Linguistics; 2002: 214-221.

88. Lee S, Lee GG: **A Bootstrapping Approach for Geographic Named Entity Annotation**. In: *Information Retrieval Technology: Asia Information Retrieval Symposium, AIRS 2004, Beijing, China, October 18-20, 2004, Revised Selected Papers.*

Edited by Myaeng S-H, Zhou M, Wong K-F, Zhang H, vol. 3411: Springer; 2005: 178-189.

89. Ciaramita M, Gangemi A, Ratsch E, Saric J, Rojas I: **Unsupervised Learning of Semantic Relations between Concepts of a Molecular Biology Ontology**. In: *Proceeding of the 19th International Joint Conference on Artificial Intelligence (IJCAI).* 2005: 659-664.

90. Agichtein E, Gravano L: **Snowball: Extracting Relations from Large Plain-Text Collections**. In: *Proceedings of the 5th ACM International Conference on Digital Libraries.* San Antonio, Texas, United States; 2000: 85-94.

91. Chaudet H: **Extending the event calculus for tracking epidemic spread**. *Artificial Intelligence in Medicine* 2006, **38**(2):137-156.

92. Schuurman I: **Spatiotemporal Annotation on Top of an Existing Treebank**. In: *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories.* Bergen, Norway; 2007: 151-162.

93. Hearst MA: **Multi-Paragraph Segmentation of Expository Text**. In: *Proceedings of the 32nd Meeting of the Association for Computational Linguistics.* Los Cruces, NM; 1994: 9-16.

94. Hearst MA: **TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages**. *Computational Linguistics* 1997, **23**(1):33-64.

95. Teufel S, Carletta J, Moens M: **An annotation scheme for discourse-level argumentation in research articles**. In: *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics.* Bergen, Norway; 1999: 110-117.

96. Teufel S, Moens M: **Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status**. *Computational Linguistics* 2002, **28**(4):409--445.

97. Mizuta Y, Korhonen A, Mullen T, Collier N: **Zone analysis in biology articles as a basis for information extraction**. *The International Journal of Medical Informatics* 2006, **75**(6):468-487.

98. Saurí R, Littman J, Knippen B, Gaizauskas R, Setzer A, Pustejovsky J: **TimeML Annotation Guidelines Version 1.2.1**. In.; 2006.

99. Partee BH: **Some Structural Analogies Between Tenses and Pronouns in English**. *Journal of Philosophy* 1973, **70**(18):601-609.

100. Bell A: **The discourse structure of news stories**. In: *Approaches to Media Discourse.* Edited by Bell A, Garrett P: Oxford: Blackwell; 1998: 64-104.

101. Schokkenbroek C: **News Stories - Structure, Time and Evaluation** *Time & Society* 1999, **8**(1):59-98.

102. Levin B: **English Verb Classes and Alternations: A Preliminary Investigation**: The University of Chicago Press; 1993.

103. Allen J: **Towards a general theory of action and time**. *Artificial Intelligence in Medicine* 1984, **23**:123-154.

104. Kawazoe A, Jin L, Shigematsu M, Barrero R, Taniguchi K, Collier N: **The development of a schema for the annotation of terms in the BioCaster disease detecting/tracking system**. In: *Proceedings of KR-MED 2006, the Second International Workshop on Formal Biomedical Knowledge Representation.* Baltimore, Maryland; 2006: 77-85.

105. Huddleston R, Pullum GK: **A Student's Introduction to English Grammar**: Cambridge University Press; 2005.

106. Palmer M, Dang HT, Fellbaum C: **Making fine-grained and coarse-grained sense distinctions, both manually and automatically**. *Natural Language Engineering* 2007, **13**(2):137-163.

107. Passonneau RJ, Habash N, Rambow O: **Inter-annotator Agreement on a Multilingual Semantic Annotation Task**. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC).* Genoa; 2006: 1951-1956.

108. Mihalcea M, Chklovski T, Kilgarriff A: **The SENSEVAL-3 English lexical sample task**. In: *Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3).* Barcelona, Spain; 2004: 25-28.

109. Bruce R, Wiebe J: **Word-sense distinguishability and inter-coder agreement**. In: *Proceedings of the Third Conference on Empirical Methods in Natural Language Processing (EMNLP-98).* Granada, Spain; 1998: 53-60.

110. Passonneau RJ, Litman DJ: **Intention-based segmentation: human reliability and correlation with linguistic cues**. In: *Proceedings of the 31st annual meeting on Association for Computational Linguistics.* Columbus, Ohio; 1993: 148 - 155.

111. Carletta J, Isard S, Doherty-Sneddon G, Isard A, Kowtko JC, Anderson AH: **The reliability of a dialogue structure coding scheme**. *Computational Linguistics* 1997, **23**(1):13 - 31.

112. Carlson L, Marcu D, Okurowski ME: **Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory**. In: *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue.* vol. 16. Aalborg, Denmark; 2001: 1-10.

113. Marcu D, Amorrortu E, Romera M: **Experiments in constructing a corpus of discourse trees**. In: *Proceedings of the ACL Workshop on Standards and Tools for Discourse Tagging.* College Park, MD; 1999: 48-57.

114. Passonneau RJ: **Computing Reliability for Coreference Annotation**. In: *Proceeding of the 4th International Conference on Language Resources and Evaluation (LREC).* Lisbon, Portugal; 2004: 1503-1506.

115. Poesio M, Artstein R: **The Reliability of Anaphoric Annotation, Reconsidered: Taking Ambiguity into Account**. In: *Proceeding of ACL Workshop on Frontiers in Corpus Annotation.* Ann Arbor; 2005: 76-83.

116. Nenkova A, Passonneau R, McKeown K: **The Pyramid Method: Incorporating Human Content Selection Variation in Summarization Evaluation**. *ACM Transactions on Speech and Language Processing (TSLP)* 2007, **4**(2).

117. Artstein R, Poesio M: **Inter-Coder Agreement for Computational Linguistics**. *Computational Linguistics* 2008, **34**(4):555-596.

118. Teufel S: **Argumentative Zoning: Information Extraction from Scientific Text**. *Ph.D. thesis.* University of Edinburgh; 1999.

119. **BioCaster text mining project.** [http://biocaster.nii.ac.jp]

120. Ramshaw L, Marcus M: **Text Chunking Using Transformation-Based Learning**. In: *Proceedings of the ACL Third Workshop on Very Large Corpora.* 1995: 82-94.

121. Charniak E: **A maximum-entropy-inspired parser**. In: *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference.* Seattle, Washington; 2000: 132-139.

122. Cohen J: **A Coefficient of Agreement for Nominal Scales**. *Educational and Psychological Measurement* 1960, **20**(1):37-46.

123. Krippendorff K: **Computing Krippendorff's Alpha-Reliability**. In.; 2007.

124. Passonneau RJ, Litman DJ: **Discourse segmentation by human and automated means**. *Computational Linguistics* 1997, **23**(1):103-139.

125. Krippendorff K: **Content Analysis: An Introduction to its Methodology**: Sage Publications, Inc; 1980.

126. Mullen T, Mizuta Y, Collier N: **A baseline feature set for learning rhetorical zones using full articles in the biomedical domain**. *SIGKDD Explorations* 2005, **7**(1):52-58.

127. Lafferty JD, McCallum A, Pereira F: **Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data**. In: *Proceedings of the Eighteenth International Conference on Machine Learning.* MA, USA: Morgan Kaufmann Publishers; 2001: 282-289.

128. **Conditional Random Fields: An Introduction** [http://www.inference.phy.cam.ac.uk/hmw26/papers/crf_intro.pdf]

129. Borthwick A, Sterling J, Agichtein E, Grishman R: **Exploiting diverse knowledge sources via maximum entropy in named entity recognition**. In: *Proceeding of the Sixth Workshop on Very Large Corpora.* Montreal, Canada; 1998: 152–160.

130. Lafferty J, McCallum A, Pereira F: **Conditional random fields: Probabilistic models for segmenting and labeling sequence data**. In: *Proceeding of the 18th International Conference on Machine Learning.* San Francisco, CA: Morgan Kaufmann Publishers Inc.; 2001: 282–289.

131. Quinlan JR: **Induction of decision trees**. *Machine learning* 1986, **1**(1):81-106.

132. Burges CJC: **A Tutorial on Support Vector Machines for Pattern Recognition**. *Data Mining and Knowledge Discovery* 1998, **2**(2):121-167.

133. **LIBSVM: a library for support vector machines** [http://www.csie.ntu.edu.tw/~cjlin/libsvm]

134. Quinlan JR: **C4.5: Programs for Machine Learning**: Morgan Kaufmann Publishers; 1993.

135. Charniak E, Johnson M: **Coarse-to-fine n-best parsing and MaxEnt discriminative reranking**. In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics.* Ann Arbor, Michigan, USA; 2005: 173-180.

136. Soon WM, Ng HT, Lim DCY: **A machine learning approach to coreference resolution of noun phrases**. *Computational Linguistics: Special issue on computational anaphora resolution* 2001, **27**(4):521-544.

137. Klein D, Manning CD: **Accurate unlexicalized parsing**. In: *Proceeding of the Association for Computational Linguistics (ACL) 2003.* Sapporo, Japan: Association for Computational Linguistics 2003: 423–430.