氏　　名	CHANLEKHA HUTCHATAI		
学位（専攻分野）	博士（情報学）		
学位記番号	総研大甲第1338号		
学位授与の日付	平成２２年３月２４日		
学位授与の要件	複合科学研究科　情報学専攻 学位規則第６条第１項該当		
学位論文題目	Document Zoning for Enhancing Spatial and Temporal Understanding in Web-based Health Surveillance Systems		
論文審査委員	主　査	准教授	COLLIER Nigel
		教授	佐藤　健
		教授	武田　英明
		准教授	古山　宣洋
		准教授	北本　朝展
		准教授	THEERAMUNKONG Thanaruk (Thammasat University)

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

Summary

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 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 an alternative means to overcome these limitations. This framework is called as spatiotemporal zoning.

The contribution of this work is to propose a scheme called spatiotemporal zoning, which analyzes each event reported in news articles with regard to its spatial and temporal information, as a 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 proposed 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 objective of the spatiotemporal zoning scheme is to enable language technology software to partition text into segments that contain group of events, which occurred in the same location within the same homogeneous time frame. More specifically, the task of spatiotemporal zoning is to classify news articles into predefined classes based on their spatial and temporal characteristics, and recognize the spatial and temporal attributes of each event. The capability of associating events reported in each text segment with the most specific spatial and temporal information available in news reports enables the systems to employ more simple techniques for detecting outbreak location. These techniques are, such as using text classification to detect text segments that indicate outbreak situations. At the same time, false alarms of past outbreaks can be avoided by taking the temporal information of events into consideration.

In news report, some text segments convey the contents that cannot be placed in time, i.e. cannot be associated with temporal information. These types of content include sentences that provide general knowledge about certain subjects, or sentences that predict or express the possibility of certain situations. The ability to distinguish event-predicates that express temporally-locatable events from other event-predicates is therefore an essential basic requirement. Apart from the spatial and temporal information, in spatiotemporal zoning scheme, news content is also classified into four classes, which are Reporting event, Normal event, Hypothetical event, and generic information.

In order to demonstrate that the proposed framework can be applied to unrestricted text, both automatically and by humans, a representative corpus was created. This corpus consists 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 employed for incorporating various sets of features in order to study the impact of textual features on event classification. In this scheme,

To recognize spatial information, several approaches, ranging from simple technique such as commonly used heuristic-based approach to a more sophisticated machine learning approach have been experimented. Various sets of features and the strategy for feature encoding were explored in order to effectively recognize spatial attribute of the events.

For temporal attribute recognition, rule-based approach is used to recognize an event's temporal information. However, it is often found that the same event is repeatedly mentioned many times, while the occurring time of the event is stated only once. 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 studies show that the proposed scheme is reliable and can be learned by human annotators. Moreover, the results from automatic zone attribute recognition show that this scheme can be done automatically with a reliable level of performance.

博士論文の審査結果の要旨

Hutchatai Chanlekha gave a 45 minute presentation in English to the examination committee which was open to the public. Dr. Thanaruk Theeramunkong from Thammasat University in Thailand – the external examiner - attended by teleconference. The major points of Hutchatai's research presentation were as follows:

- Unstructured textual resources on the Web such as media documents are becoming widely used as a source for health surveillance. In order to make the best use of these resources human analysts as a minimum need to know the time and place where events of interest take place. Current technology for automatically detecting geo-temporal information appears inadequate for the practical realization of this goal. In particular the methods are not adequate to differentiate all times/locations mentioned in the text or to find information at the finest levels of granularity.
- Areas surveyed include extant temporal schemes such as TimeML and Chaudet's SpatioTemporal Extended Event Language (STEEL) as well as operational systems such as HealthMap, GPHIN, BioCaster and MedISys.
- Based on an extension of the ideas in rhetorical zone analysis put forward by Teufel and Moens, Hutchatai contributes to the task in two areas: (1) in knowledge representation by developing a novel scheme for annotating textual areas according to regions with homogeneous geo-temporal characteristics and testing its reproducibility on tests with human annotators, and (2) in knowledge acquisition to show how the scheme can be automatically annotated using supervised machine learning.
- The presentation outlined the geo-temporal zone scheme and experiments were conducted to show the level of human inter-annotator agreement. Results show a high degree of agreement between annotators (about 0.9 kappa score).
- The resulting gold standard data was then used in machine learning experiments to show automatic annotation.
- In automatic analysis a range of linguistic features are compared to classify the event type (normal, reporting, information, hypothetical) for each clause and the results are compared. For spatial recognition several machine learning models (conditional random fields, support vector machines, C4.5 decision trees) were compared across various linguistic features to show those that performed the best. Qualitative analysis is then performed on the error cases to highlight key areas of confusion, e.g. between normal and information events, by propagating errors between adjacent clauses, and where events start and end at different locations.

There followed an open question and answer session at which the committee members asked a number of questions:

- The exact role of 'normal' events compared to 'information' events was clarified.
- The application of the technique to multiple documents as future work was discussed.

- The extensibility of the method was discussed to events outside of health surveillance.
- The committee requested a clarification of the assumption that the four event types were mutually exclusive to be included in the thesis.
- The committee members recommended that the conclusion should contain a statement detailing how the reported method could be fitted into a practical health surveillance system, i.e. how much work would be required to make it operational.
- The final version of the thesis also needs to briefly comment on the degree of domain and language dependency of the reported techniques.

The committee members expressed their satisfaction with the quality of the answers, the work reported in the thesis, its novelty and the overall approach. The committee recognized the original contribution made by the research in the area of knowledge representation and acquisition as well as the achievements reported in the international journal publication in BMC Medical Informatics and Decision Making.

The committee members all agreed that Hutchatai Chanlekha had successfully passed the thesis examination.