Text Information Extraction and Aggregation in a Mobile-based Emergency Response System

Khaled Amailef, Jie Lu, Jun MA

Faculty of Engineering and Information Technology
University of Technology Sydney, Po Box 123 Broadway
NSW, 2007, Australia
email: kamailef@it.uts.edu.au, jielu@it.uts.edu.au, junm@it.uts.edu.au

Abstract

A mobile-based emergency response system (MERS), as one of the important Mobile Government (m-Government) services, aims to reduce risks in an emergency situation. This paper describes an algorithm within MERS applications to automatically extract information from SMS data based on an ontology concept, a maximum entropy statistical model, and a set of fuzzy rules. The algorithm has four main functions: collect unstructured information from Short Message Service (SMS) emergency text message; conduct information extraction and aggregation including lexical analysis, name entity recognition, merging structure, and normalization and duplication; calculate similarity of SMS text messages; and generate query and results presentation.

Keywords: m-Government, mobile-based systems, emergency response systems, information extraction, fuzzy rules

1. Introduction

Within the context of electronic government (e-Government), m-Government services offer more access to information and services for citizens, businesses, and government. Literature shows that mobile-based information systems can be a solution to benefit first responders in a variety of ways. First, the mobility devices allow first responders to perform better rapid and actionable decision making. Second, the development of mobile technology and communication trends provides an ubiquities environment for deployment within a variety of fields [1]. We have developed a mobile-based emergency response system (MERS) [2], aims to makes the use of mobile technologies to assist government to get information and make decisions in response disasters. The MERS supports five major applications on emergency response including registration, monitoring, analysis, decision support and warning generation.

Two of the most important tasks within analysis applications are information extraction (IE) and aggregation. IE is 'a technology for finding facts in plain text, and coding them in a logical representation such as a relational database' [3], and aggregation aims to combine two or more linguistic structures into a single sentence

In this paper, we present an algorithm to automatically extract information in an emergency situation from SMS text messages. The algorithm based on ontology concept for unstructured information, a maximum entropy statistical model with various predefined features and fuzzy rules.

This paper is structured as follows: Section 2 provides related work in emergency response system and text extraction concept. Section 3 describes an information extraction and text aggregation algorithm in MERS. Finally, conclusions and future work are highlighted in Section 4.

2. Related Works

For the convenience of describing proposed algorithm, we will first introduce a briefly description of MERS, and then, we will give some related concept of information extraction which will be used in the following sections.

2.1. Mobile-Based Emergency Response System

This section provides the concept of a MERS system under an m-Government platform. The structure of the system developed includes three main parts: m-Government dimensions, MERS project and end-users. The MERS project within this study consists of four main components: inputs, processes, outputs, and outcomes:

- **Inputs:** elements to enter into the system. Examples of inputs are a mobile user's details and emergency request data.
- **Processes:** all the elements necessary to convert or transform inputs to outputs. For example, a process may include disaster monitoring, and data analyzing.
- **Outputs:** the consequences of being in the system. For example, warning messages.
- **Outcomes:** can take any or both of two forms benefits and/or risks, each should be well planned.

2.2. Named Entity Recognition

Named entity recognition (NER) is important for semantically oriented retrieval tasks, such as question answering, entity retrieval, biomedical retrieval, trend detection, and event and entity tracking [4]. Much research has since been carried out on NER, using both knowledge engineering and machine learning approaches [5].

2.3. Maximum Entropy Model

Information extraction can be viewed from a statistical point of view. The maximum entropy (ME) model has been applied in many areas of science and technology such as natural languages processing, text classification, and machine learning [6]. The ME is a flexible statistical model which assigns an outcome for each token (word) based on its history and features. We can compute the conditional probability by:

$$p(o \setminus h) = \frac{1}{Z(h)} \prod_{j=1}^{k} \alpha_j^{f_j(o,h)} \quad (1)$$

$$Z(h) = \sum_{h} \prod_{j=1}^{k} \alpha_j^{f_j(o,h)} \quad (2)$$

$$t = \arg\max p_j(o \setminus h) \quad (3)$$

Where o refers to the outcomes and h is the history or the context, history can be viewed as all information derivable from the training corpus relative to the current token [7]. $Z(h)$ is a normalization function and is equal to 1 for all h. $\{f_1, f_2, \ldots f_k\}$ are known as features. Parameter α_j can be calculated by use of Improved Iterative Scaling (IIS) algorithm [8], note that each α_j corresponds to f_j, where $f_j(o,h) \rightarrow \{0,1\}$, nnd t is the optimal probability distribution according to the maximum likelihood criterion.

2.4. Word Feature Generation and Selection

In this section, we describe the features used in MBERS. Since the features are critical to the success of machine learning approaches [9], we will discuss them in more detailed. The features used in our model come from reference [10]:

Slipping window: typically, word windows are of size 5 consisting of the current token, two tokens to left and two tokens to right (e.g. w_{-2} w_{-1} w_0 w_1 w_2). This feature is useful for determining the class to which entity the current token belongs.

Previous token tags: These features are helpful to capture some information on interdependency of the label sequence.

Prefix and suffix: These features provide the composition of the word, which usually yields useful information on its semantic.

Capitulation and digit information: Semantic or syntactic type information can provide useful information.

Dictionary lookup features: This feature is used to decide whether a token phrase belongs to a certain prebuilt dictionary or not.

Pattern feature: This feature is used to determine a set of decision rules that would produce a positive example.

3. An Information Extraction and text Aggregation Algorithm

In this section we present the algorithm of the information extraction and SMS text aggregation. It is used to automatically extract structure information from unstructured SMS information.

3.1. The Components of SMS text Extraction and Aggregation Process

Our overall goal is to extract and aggregate SMS text messages received from mobile phone users in emergency situation. To reach this goal we are applying several processing steps. First, unstructured information is pre-processed by storing into a relational database. Then, information extraction algorithm is applied. The result of this information extraction algorithm is stored in a relational database. For the decision makers, the structured information in term of tables and messages can be used to assist in responding to an emergency situation. The main components of the SMS text extraction and aggregation architecture and their interactions are depicted in Fig 1.

3.2. Algorithm

In the following, we propose an algorithm for SMS text extraction and aggregation.

Input: Emergency SMS text messages;
Output: SMS text classification
Purpose: First, to identify all pre-defined of the structured information in the unstructured text. Second, the goal is to populate a database of structured entities.
Step 1: collect unstructured information of each individual SMS emergency text message.
The tasks of this step are to lookup resources. A collection of SMS emergency text message is acquired and depositing them in relational database for further analysed.
Step 2: conduct information extraction and aggregation.
Information extraction is the task of finding specific pieces of information from unstructured or semi-structured document. The Information extraction algorithm includes three step showed as following:
Step 2.1: lexical analysis.
The SMS text is first divided into tokens, and uses stopping technique to remove information that is not useful for categorization (so-called stop-words) such as articles, propositions, pronouns and other functional words that are not related to the content.

Step 2.2: Name entity recognition.
Once a name entity (NE) is identified, it is then classified into predefined categories to SMS text messages such as location, type of disaster, physical target etc. We use a maximum entropy statistical model with various predefined features for named entity recognition.
Step 2.3: Merging structure.
This step is referred to as relation extraction for the case in which two entities are being associated.

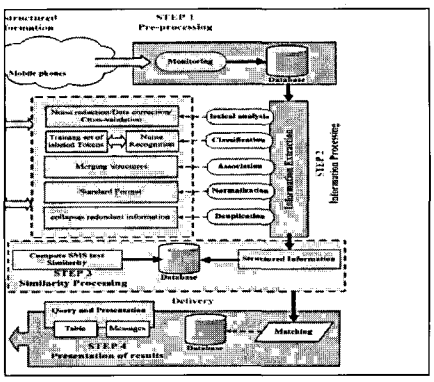

Fig. 1: Architecture for SMS text extraction and aggregation to a data source.

Step 2.4: Normalization and duplication.

Puts information in a standard format in which it can be reliably compared, and avoid to get redundant information in our database.

Step 3 Calculate similarities of composite SMS text messages

The similarity approach is based on overall sentence similarity [11]. The method provides a similarity score between 0 and 1. If the similarity score is above a certain threshold then the elements are considered as match candidates. We construct a string matching matrix M and repeat the finding of the maximum-valued matrix-element and removing all the matrix element of the corresponding row and column until matrix M is empty. This method is modified based on a set of fuzzy rules [12] as illustrated in Fig. 2, so that each maximum-value matrix-element is equivalent with the fuzzy term set:

$Very\ Low(VL) = \{(\mu_{VL}(x), x) | x \in X\}$,

$Low(L) = \{(\mu_{L}(x), x) | x \in X\}$,

$Medium(M) = \{(\mu_{M}(x), x) | x \in X\}$,

$High(H) = \{(\mu_{H}(x), x) | x \in X\}$, and

$Very\ High(VH) = \{(\mu_{VH}(x), x) | x \in X\}$ is obtained based on equations (5), (6),(7),(8) and (9), these elements have value of ranging from 0 (not similar) to 1 (maximum similarity).

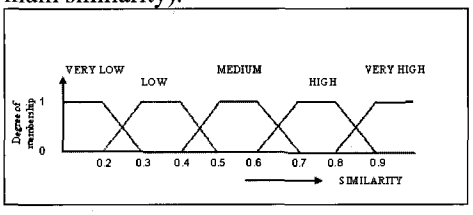

Fig. 2: Fuzzy sets to characterize the SMS text similarity.

$$\mu_{VL}(x) = \begin{cases} \frac{x-0.2}{0.1}, & for\ 0.2 \le x < 0.3 \\ 1 & for\ 0.0 \le x \le 0.2 \\ 0 & for\ x > 0.3 \end{cases} \quad (4) \qquad \mu_{L}(x) = \begin{cases} \frac{x-0.2}{0.1}, & for\ 0.2 \le x < 0.3 \\ 1 & for\ 0.3 \le x \le 0.4 \\ \frac{0.5-x}{0.1}, & for\ 0.4 < x \le 0.5 \\ 0 & for\ x > 0.5 \end{cases} \quad (5)$$

$$\mu_{M}(x) = \begin{cases} \frac{x-0.4}{0.1}, & for\ 0.4 \le x < 0.5 \\ 1 & for\ 0.5 \le x \le 0.6 \\ \frac{0.7-x}{0.1}, & for\ 0.6 < x \le 0.7 \\ 0 & for\ x > 0.7 \end{cases} \quad (6) \qquad \mu_{H}(x) = \begin{cases} \frac{x-0.6}{0.1}, & for\ 0.6 \le x < 0.7 \\ 1 & for\ 0.7 \le x \le 0.8 \\ \frac{0.9-x}{0.1}, & for\ 0.8 < x \le 0.9 \\ 0 & for\ x > 0.9 \end{cases} \quad (7)$$

$$\mu_{VH}(x) = \begin{cases} \frac{x-0.8}{0.1}, & for\ 0.8 \le x < 0.9 \\ 1 & for\ 0.9 \le x \le 1.0 \\ 0 & for\ x > 1.0 \end{cases} \quad (8)$$

Then we conduct linguistic aggregation operation as follows:

$S\alpha, S\beta \in \bar{S}, \lambda \in [0,1]$, Then their operational laws can be defined as follows:

$S_{\alpha} \oplus S_{\beta} = S_{\alpha+\beta}$, $\lambda S_{\alpha} = S_{\lambda\alpha}$, where

$\bar{S} = \{S_{-2} = very\ low, S_{-1} = low, S_{0} = medium, S_{1} = high, S_{2} = very\ high\}$

The two SMS text messages are highly similar, if the balance similarity score that is obtained from linguistic aggregation is high. If the two sentences are similar, we compute the union operation between them, so, $U=P \cup R$.

Step 4: Generate queries and results

Provide the decision support makers and mobile users with query related to emergency situation. A simply query, for example, would allow mobile phone users to receive a warning emergency text messages. A query also allows decision support makers to analysis the emergency current situation.

3.3. Ontology-based Representation for Unstructured SMS Text

Ontology has been used in many information extraction systems [13]. Ontology can be defined as knowledge expression that allows us to share understanding of some domains of interest. Ontology defines the concepts and various relations among the concepts [14].

In this paper, ontology for the SMS text messages domain has been developed for experimental purposes of text classification. In the domain ontology, we construct a SMS text message domain ontology which consists of a seven main entities including: disaster location; stage of execution; human target; instrument used; physical target; disaster event; and date and time using Protégé tool. All entities are dividing into categories and subcategories.

Our goal is to extract key words in a SMS emergency text message that bear most of useful information content of the received SMS text message. We defined the following classes of words in a SMS emergency text message as illustrated in Fig 3:

- **DisasterLocation:** This class represents any location in SMS message that connected to the disaster event. For example, State, City, and Street name.
- **StageOfExecusion:** This class represents the status of disaster. For example, Attempting, in progress or accomplished.
- **HumanTarget:** This class represents the number of injured and death.
- **InstrumentUsed:** This class represents the instrument that used in the disaster event such as a vehicle bomb.
- **PhysicalTarget:** any building type mentioned in SMS text is a physical target such as educational building.

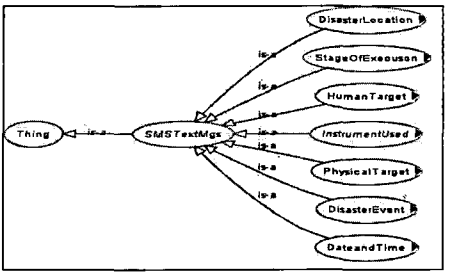

Fig. 3: Ontology-based concept hierarchy for SMS Text Extraction and aggregation.

- **DisasterEvent:** This is the class of disaster event. For example, fire and terrorist attack.
- **DateandTime:** This class represents the date and the time of disaster event.

4. Conclusions and Future Work

In this paper, we have presented an algorithm which able to extract and aggregate information from SMS text messages as a part of MERS applications. First, we introduced the components of information extraction and aggregation process. The proposed algorithm can extract many kinds of semantic element of the emergency situation information such as disaster location, disaster event, status of disaster, etc. The algorithm also used a maximum entropy statistical model for name-entity recognition and calculated message similarity based on fuzzy rules.

The proposed algorithm is being implemented in a software prototype for emergency information aggregation and disaster management. Further study will include the development of approaches for more complex problems with SMS text automated extraction using the software for MERS.

ACKNOWLEDGMENT

The work presented in this paper was supported by Australian Research Coun-

cil (ARC) under ""Discovery Project DP0880739.

References

[1] S. Kim, *et al.*, "Mobile analytics for emergency response and training," *Information Visualization, Houndmills*, pp. 77-88, 2008.

[2] K. Amailef, and J. Lu, "m-Government: A Framework of Mobile-based Emergency Response Systems," *International Conference on Intelligent System and Knowledge Engineering (ISKE2008)*, 2008.

[3] Y. Roman, *et al.*, "Extracting information about outbreaks of infectious epidemics," *Proceedings of HLT/EMNLP on Interactive Demonstrations*, 2005.

[4] J. Valentin, *et al.*, "Named entity normalization in user generated content," *Proceedings of the second workshop on Analytics for noisy unstructured text data*, 2008.

[5] C. Hai Leong and N. Hwee Tou, "Named entity recognition with a maximum entropy approach," *Proceedings of the seventh conference on Natural language learning at HLT-NAACL*, 2003.

[6] A. A. Javed, *et al.*, "The maximum entropy method for analyzing retrieval measures," *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 2005.

[7] S. S. Kumar, *et al.*, "Feature selection techniques for maximum entropy based biomedical named entity recognition," *Journal of Biomedical Informatics*.

[8] G. Liang, *et al.*, "Maximum entropy models: convergence rates and applications in dynamic system monitoring," *Information Theory*, pp. 168, 2004.

[9] C.-W. WU, *et al.*, "Learning to Integrate Web Taxonomies with Fine-Grained Relations: A Case Study Using Maximum Entropy Model," *Second Asia Information Retrieval Symposium* , pp. 190-205, 2005.

[10] S. Weiss, *et al.*, *Text Mining: Predictive Methods for Analyzing Unstructured Information*, 2005.

[11] I. Aminul and I. Diana, "Semantic text similarity using corpus-based word similarity and string similarity," *ACM Trans.* pp. 1-25, 2008.

[12] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 338-353, 1965.

[13] D. Tran Quoc and W. Kameyama, "A Proposal of Ontology-based Health Care Information Extraction System: VnHIES," *Innovation and Vision for the Future, IEEE International Conference*, pp. 1-7, 2007.

[14] Z. LI and K. Ramani, "Ontology-based design information extraction and retrival," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, pp. 137-154, 2007.