Automatic Semantic Description Extraction from Social Big Data for Emergency Management

Bukhoree Sahoh, Anant Choksuriwong

Department of Computer Engineering, Faculty of Engineering, Prince of Songkla University, Songkla, Thailand
sahoh.b@gmail.com, anant.c@psu.ac.th (✉)

Abstract. Emergency events are unexpected and dangerous situations which the authorities must manage and respond to as quickly as possible. The main objectives of emergency management are to provide human safety and security, and Social Big Data (SBD) offers an important information source, created directly from eyewitness reports, to assist with these issues. However, the manual extraction of hidden meaning from SBD is both time-consuming and labor-intensive, which are major drawbacks for a process that needs accurate information to be produced in real-time. The solution is an automatic approach to knowledge discovery, and we propose a semantic description technique based on the use of triple store indexing for named entity recognition and relation extraction. Our technique can discover hidden SBD information more effectively than traditional approaches, and can be used for intelligent emergency management.

Keywords: Ontology, natural language processing, information extraction, semantic index, named entity recognition, triplestore

1. Introduction

Emergency events are unpredictable and undesirable, be man-made (e.g. terrorism, transportation accidents) or natural disasters (e.g. floods, earthquakes), and affect lifestyle and government infrastructure. An emergency authority must have access to real-time information to deal with these events successfully (Sahoh and Choksuriwong 2018).

In the past two decades, Social Big Data (SBD) has played an important role as a real-time data source in the emergency field (Castillo 2016, Sahoh and Choksuriwong 2017). SBD is focused on data from social media that the big data is massive, heterogeneous, and streamed in real-time. It offers a flexible perspective, based on a collaborative vision, to explain an event from different points of view (Bello-Organ et al. 2016). This can help the emergency authorities understand events deeply, and guide their decision-making. This is especially true of emergency SBD where natural language text can vary in word choice, morphology, and syntax during stressful situations. Also, the text consists of both explicit

information, directly expressed in the SBD content, and implicit information that must be inferred. This means that although SBD is meaningfully understood by emergency authorities, SBD understanding is labor-intensive and time-consuming. In addition, the effectiveness of emergency management depends on information being available promptly. This shows a need for both automatic and real-time processing to deal with SBD event understanding.

The automatic requirement involves making machines understand the meaning of textual SBD, which necessitates a semantic description. This will identify the meaning of unstructured data and transform it into emergency information in a machine-readable format. Two approaches for the semantic description of emergency SBD are Named Entity Recognition (NER) and Relation Extraction (RE). They recognize SBD transactions and their relationships, to reveal both explicit and implicit information for later analysis. Highly accurate semantic descriptions effectively support automatic SBD analysis for

decision-making.

Traditional techniques for natural language understanding are word matching and semantic analysis based on word order. Word matching is utilized by Emergency NER (ENER) to name an entity given in the textual SBD by using a dictionary-based emergency corpus to describe the event (Choi and Bae 2015, Sandhu et al. 2016). This technique uses a word-to-word similarity measure that requires the prior disambiguation of word senses, which is not possible for the natural and informal language in SBD. Instead, some studies have proposed using a semantic analysis based on word orders by applying pattern-based matching to annotate the textual data (Oramas et al. 2016, Velasco-Elizondo et al. 2016, Zhou and El-Gohary 2017). For example, this disambiguates words by considering word order as a context. However, this technique cannot understand implicit SBD semantics to uncover hidden meaning. By hidden meaning, we mean an implicitly named entity and its relationship, discovered by considering the semantic context using human-like inference. This includes the set of word orders and syntactic structures used as a bridge between disconnected explicitly named entities. A lack of semantic understanding is one cause for the high recall but low precision in ENER, which causes irrelevant entities to be used to describe an emergency event. Inaccurate information of this kind can cause the emergency authorities to misunderstand the situation, make wrong decisions, and thereby make the situation worse.

To deal with these problems, we propose a novel automatic semantic description technology for dealing with SBD. In particular, it focuses on a semantic-based bridge for uncovering the hidden SBD meaning by considering the deep meaning of entities through their context in an emergency. We utilize triple store indexing for semantic context analysis to name relevant emergency entities and their relationships, which lets the emergency event be de-

scribed more clearly and meaningfully. Our results show that the triple store index can provide higher accuracy than traditional techniques for the ENER process. Therefore, the semantic description of an emergency event can be used to support better decision-making in emergency management.

2. Related Work and Background Knowledge

This section considers event descriptions based on natural language understanding and reviews the current approaches for discovering the emergency events. An ontology-based model for emergency event understanding is investigated to discuss the challenges of applying this approach to automatic event description.

2.1 An Existing Event Description for Emergency Event Understanding

Emergency event understanding analyzes big data to discover knowledge for supporting human decision making (Shah and Zimmermann 2017). The aim is to identify key features from a real-world environment, such as SBD, and employ them to describe that event. Event description is an essential process in emergency event understanding, especially for SBD (Murphy 2016, Riccardi 2016). SBD emergency information can be divided into explicit and implicit information. Explicit information defines the entities that are clearly represented by word choice while implicit emergency information comes from the hidden meaning of entities, discovered by bridging explicit information, and inferring it from prior knowledge. ENER is a key method for event description, by recognizing meaningful entities from the SBD.

ENER for emergency management has been proposed for dictionary-based word matching (Jing et al. 2016, Rossi et al. 2018), and cosine-similarity word matching (Farag et al. 2018). Some studies have employed rule-based and

dictionary-based word matching to describe an SBD emergency event (Xu et al. 2016, Yang et al. 2018), but without implicit emergency information inferred from natural language. This paper addresses that limitation by considering hidden meaning using semantic context, by applying a semantic analysis to ENER, also known as a semantic description.

Semantic description is a natural language processing that determines the meaning of an entity (linguistic unit) by considering its semantic context. The semantic context of an entity is the word ordering which shows the dependency relations between the entities. It can be used to bridge disconnected explicit entities to reveal the implicit entities (Derczynski et al. 2015). In this paper, the semantic context has two main attributes: (1) a semantic-based language dependence for describing an instance, and (2) a semantic-based bridge for describing the context of those instances. For example, the phrase "*The emporium shopping mall*", reveals its semantic-based language dependence by using "*shopping mall*" as word order. This lets "*The emporium*" be explicitly annotated as an instance of a "*Marketplace*".

An earlier study employed automatic semantic description using semantic-based language dependences (Piao et al. 2019). An entity was observed by matching word-ordered patterns with semantic categories, or parts of speech, to describe the entity's meaning explicitly. Another study focused on negation patterns by considering negative-prefix word dependences to define an adverse entity meaning (Jiménez-Zafra et al. 2018). However, both those works only looked at word identification based on lightweight semantic concepts. They did not consider semantic-based bridges for identifying semantic relations in real-world events. This is a significant problem in SBD, which primarily represents implied content. Also, some implicit information needs background knowledge to support its inference, which can be addressed by combining the semantic-based bridge with ontology models

(Stepanov et al. 2018, Stork et al. 2018). This can describe both directly and indirectly related entities by identifying hidden relations. Returning to the earlier example, a semantic-based bridge can discover the indirect relation of "*Marketplace*" to "*Bangkok*" as a "*Location*" concept. Moreover, "*Marketplace*" can be ridged to "*Customers*" as a "*Citizen*" concept. Unfortunately, these approaches need the addition of events of interest, known as semi-automatic descriptions, which is both labor-intensive and time-consuming.

The automatic semantic description of SBD has two main requirements: (1) An ontology-based model for the emergency event to represent the semantic context for both the semantic-based language dependencies and the semantic-based bridge, and (2) a fully automatic semantic description approach based on the real-time processing of the massive volume of SBD. These requirements are the principal elements of emergency management with well-timed information. An ontology-based model for emergency event understanding is clearly a key requirement for supporting automatic semantic descriptions.

2.2 An Ontology-Based Model for Emergency Event Understanding

An ontology-based model is required for dealing with SBD characteristics (Kuflik et al. 2017, Poblet et al. 2018) because it can represent emergency knowledge as concrete-semantic concepts. An ontology-based model with "*Subject*" and "*Object*" as classes, and "*Predicate*" as a relationship between them, is called a triple store (Pham et al. 2018). "*Subject*" is a "*Domain*" constraint and "*Object*" a "*Range*" constraint which can represent a semantic bridge among classes. For example, when a class is recognized, then its related "*Domain*" and "*Range*" values are automatically bridged by identifying the relationships in the semantic context. "*Subject*" and "*Object*" properties can be utilized for finding language-based dependencies and "*Predicate*" properties are employed for creating

semantic-based bridges.

In emergency management, ontology can support the authorities when responding to emergency events by interpreting emergency information reasonably. Li et al. (2019) proposed an ontology for clustering emergency scenarios which can be employed for understanding the occurrences of the emergent events. Meng et al. (2018) present an ontology for the emergency response system, while Bitencourt et al. (2018) contributed information integration between authorities and systems.

Ontology can act as ground truth for the emergency event model utilized in the emergency management system (Fernández et al. 2010). However, current research is aimed at developing ontologies for supporting emergency technicians rather than providing automatic semantic descriptions for SBD. The goal of event descriptions in this research is to apply existing ontology-based emergency models and triple stores to the automatic semantic description. Triple stores can also be indexed so they can be utilized in both language-based dependencies and semantic-based bridges for the automatic semantic description of SBD.

3. An Overview of the Architecture for Automatic Semantic Descriptions

Automatic semantic description for emergency understanding is a real-time process that transforms textually unstructured data into a machine-readable form. It is based on principal SBD analytics (Guo et al. 2017, Murthy and Gross 2017, Olshannikova et al. 2017), consisting of five components: (1) Data streaming, (2) Tokenization and noise removing, (3) Emergency named entity recognition, (4) Emergency event description, and (5) Event information creation.

According to Figure 1, real-time data streaming collects emergency reports as raw textual SBD, and passes it to tokenization and noise removing. Emergency dictionary-based tokenization is applied to extract words from the raw text as tokens. Due to the nature of

informal writing in SBD, noise removal is necessary to filter out irrelevant entities. The results entities are then annotated by defining relationships between them.

ENER annotates explicit token entities using the RexEu and triple store indices. RexEu is a semantic-based language dependence approach for emergency event understanding which employs patterns to recognize data such as affected persons, and the incident times and locations in groups of conditional entities. A triple store index supports a semantic-based bridge by annotating implicit token entities. In SBD, these are irregular pattern tokens or specific words, such as emergency event names, related objects, or effects. ENER outputs named entities that represent the context of the emergency event as meta-information according to the semantic knowledge base. In emergency event description components, the emergency-named entities and their meta-information will define their relationships as statements using ontology. This employs semantic structures and constraints that represents the relation between entities in terms of subject, predicate, and object. The emergency information creation component applies a semantic parser to transform the statements into emergency information, and stores them in an emergency knowledge base.

A triple store index for semantic descriptions requires semantically graphical knowledge, or an ontology, as a triple store, based on the emergency domain. A semantic parser is employed to transform the triple store into a triple store index. The resulting index can be used to automatically match an entity from the SBD with both ontology structures and vocabularies, as discussed in more detail in the next section.

4. Triple Store Index

Emergency SBD emerges with a context description in the form of word orders, attributive adjectives, attributive nouns, and attributive verbs. There is also additional textual data

Figure 1 The Overview Architecture for Automatic Semantic Descriptions

in hidden relations that only humans can understand. For example, every emergency event happens at some location at some time, which means that the event is intuitively related to both concepts.

Accordingly, the machine should imitate human senses by automatically extracting key entities and relationships which bridge the raw data and the emergency information.

The major goal of a triple store index is to hold the knowledge that the machine can automatically use to interpret and understand critical events. Our work contributes a triple store index that gives meaning to SBD for emergency information. It does this automatically by building a semantic description by transforming emergency-unstructured data into a machine-readable form.

Triple store index construction consists of components for indexing and searching. These must also be an ontology for the emergency domain, which is used to index the triple store.

4.1 Emergency Ontology for Automatic Semantic Descriptions

Our ontology-based model for emergency understanding extends an existing model for on-

line news descriptions (Xu et al. 2016). Most of the emergency SBD is provided by crowd-sourcing, who are non-professional reporters who supply data in a short, informal manner. The main contributions of our ontology remodeling are: (1) the determination of the key concepts and their relationships for emergency events according to the needs of a semantic-based bridge, and (2) a redesign of the ontology properties, such as new labels and classes. This was done by determining the proper nouns, abbreviation taxonomies, informal language terms, and synonym terms necessary to make the triple store index more flexible based on the SBD's natural language.

Every entity or resource in the ontology model has its own properties which are represented principally in the triple store. The triple store uses the semantic pattern [entity class]-[bridge/link]-[entity class], which lets each entity class bridge to multiple classes via multiple relationships. This meta-data form can be searched and interpreted semantically. An example of this form of ontology remodeling is shown in Figure 2.

There are six classes used in Figure 2: "Event", "Person", "Location", "Timestamp", "Effect", and "Related_Object". The "Event"

Figure 2 The Extended Emergency Case Ontology for SBD

class can describe crisis circumstances such as "shooting", "bombing", "sabotage", and "riot", and refers to a critical situation which needs essential information for assessing the event. This is achieved by bridging the event with the context as relationships between classes. For example, to understand the emergency "Event" requires "Person", "Location", and "Timestamp". "Person" depicts a stakeholder in the emergency event, including a person in authority, a citizen, or a criminal. The impact of a "Person" is described by "Effect" which may refer to death, minor injury, or serious injury, which will require prioritization for different types of help. "Location" is linked to the place where the event occurred. It is important that the place can be reached quickly by emergency services, and so is detailed through multiple sub-area hierarchies that provide specific details (In the case of Thailand, Location is divided into Province, District, Sub-district, and Village). "Timestamp" explains the incident period, including festival season, weekend, holiday, and rush hour, which will affect planning

differently.

This approach describes an event using a triple store semantic pattern, such as [Event class]-[hasAffectedPerson]-[Person class] and [Person class]-[hasEffect]-[Effect class]. These triple stores will be used to automatically describe the emergency event from SBD by indexing.

4.2 Triple Store Indexing

Indexing reuses existing triple stores and applies them to describing an SBD emergency event automatically. Indexing creates an index document from the semantic pattern, called a Triple Store Index Document (TSID). TSIDs contain a triple store employing eXtensible Markup Language (XML), Resource Description Framework (RDF), Resource Description Framework Schema (RDFs), or Web Ontology Language (OWL). A triple store based on OWL consists of emergency class entities and their relationships, which are represented by a subject, predicate, and utilize constraints on their domain and range. According to our emergency ontology, TSIDs store three elements ac-

cording to the semantic pattern [entity class as **rdfs:domain**]-[bridge/link as **owl:property**]-[entity class as **rdfs:range**]. Examples are given in Table 1.

TSIDs for emergency event information are shown in rows 1-4 of Table 1. From row 4, the "#Event" resource is a **subject** with a "#happenIn" resource as its **predicate** that links to a "#Location" resource as an **object**. In the sixth row, the "#Province_instance" resource has relation "#rdfs:subClassOf" with the "#Location" resource, which means that the province is where the emergency event occurred.

The triple store is utilized by ENER for searching for meaning in multiple properties. Suppose that "Province_instance" is a resource that represents province elements from the triple store, then it can annotate an observed entity with matched TSIDs since "#province-Name" can represent the "Province's named entity" in different languages, abbreviations, or by using informal names in local languages. These TSIDs will be stored in the triple store index directory, and used to name the instance-related entity based on triple store index searching.

4.3 Triple Store Index Searching

In an emergency, textual data from formal documents (such as government reports) and informal documents (such as SBD reports) are all represented in natural language.

Triple store index searching matches those terms with all the TSIDs in the triple store index directory by applying the well-known semantic search algorithm based on domain and range identification to bridge the emergency entity (Sayed and Al Muqrishi 2017). Domain and range identification identifies the context of the searching terms according to the semantic pattern. TSIDs, a triple concept, can be used to search the semantically relevant documents with emergency terms by inferring from their contexts automatically. This will be aided by applying the triple store indexing of the ontology-based emergency model.

The set of entities coming from SBD are

streamed, and may consist of "event entities", and related "object instances" such as "province names", "affected first person names", and "timestamps". Those entities will be sent to the triple store index searching component, where the domain and range with other entities will be identified from knowledge bases stored as TSIDs. In other words, a process of semantic-based bridge-building is being added in order to identify hidden relations. Figure 3 shows a simple mechanism for triple store index searching based on this example.

Search consists of the three components shown in Figure 3; emergency-related entities, triple store index searching, and the triple store index directory. The table represents the semantic meaning output for the emergency-related entities and the triple store index. The emergency-related entities are keyword terms that were streamed from the SBD transaction. For example, a search of the triple store index for the token entities "Event Entity" and "Province Name" matches against TSIDs with the properties "#Event_Instance" and "#Province_Instance" according to the second and fourth rows respectively. This means that the ENER process can recognize both entities and use them to extract the relationship "#happenIn" from row seven. This describes how a "#Event_instance" can happen in "#Province_instance", as shown in the fifth row, which helps the machine understand the meaning more intelligently.

Triple store index searching requires functionally that can match an entity from SBD in the context of semantic-based language dependencies. RexEu offers a powerful way to identify a generalized context pattern from textual SBD. The goal of RexEu is to define emergency templates of identical elements such as affected persons, sensitive areas, and surveillant infrastructures. RexEu examples are given below.

According to RexEu Affected Person, a group of entities always comes together with the person's title, first name, and family-name in dependent order, which are the context for

Table 1 Examples of TSIDs for Triple Store Indexing

No.	Subject (rdfs:domain)	Predicate (owl:property)	Object (rdfs:range)
1.	#Event	#hasAffectedPerson	#Person
2.	#Person	#hasEffect	#Effect
3.	#Event	#happenAt	#Timestamp
4.	#Event	#happenIn	#Location
5.	#Province	RDFS:subClassOf	#Location
6.	#Province_instance	RDF:type	#Province
7.	#Province_instance	#provinceName	'Province's named entity'

represents a resource and '_' represents a string or literal.

Figure 3 Triple Store Index Searching

each other. This is a vital process for semantic-based language dependencies. Especially for emergency reports, the affected person can be either a citizen or an officer, and has the same pattern of personal identity except for their title which refers to the person type in different contexts.

Semantic-based language dependencies function as an initial filter to screen emergency events instances before their meanings and relationships are determined. It can support triple store index searching, which increases ENER effectiveness.

From this viewpoint, the triple store index is used for automatic semantic description in an emergency management system. It helps non-semantic-expert users by just inputting the knowledge graph, which is used to construct the triple store index automatically. Moreover, the SBD entity can be searched automatically and be sensibly named, based on its context. A triple store index is useful for the named entity recognition component, and for event description used by SBD analysis for the automatic semantic description.

RexEu Affected Person = [('Title')('First name')(Space) ('Family name')]

Citizen Title ::= ('Mr.' | 'Mrs.' | 'Ms.')

Officer Title ::= ('Sergeant:Sgt.' | 'Corporal: Cpl.' | 'Captain:Capt.')

RexEu Location = [('Location type')('Space') ('Location name')]

AreaType ::= ('Province' | 'district' | 'subdistrict' | 'village')

PlaceType ::= ('School' | 'Mosque' | 'Temple' | 'Hotel')

5. Case Study: Automatic Semantic Description

This study proposes the use of automatic semantic descriptions for discovering useful information in the SBD. It is especially applicable to emergency management, which needs its ad-hoc and precise ability for emergency event discovery, as this section will show. We present an automatic semantic description application for discovering emergency information from SBD for Southern Thailand terrorism. The raw data are popular news reports in Thai, crowdsourced via Twitter. Twitter's limitation of 140-character messages can be seen as an advantage since a user will include important information necessary for automatic semantic descriptions.

To remodel ontology, 10,000 tweets were accumulated between February 2 and May 9, 2018, randomly sampled from emergency environments and manually grouped by principal terms for a design class, instant, and property. Basic emergency terms are utilized such as proper nouns, abbreviation taxonomies, informal language terms, and synonym terms, determined by experts and practitioners from DeepSouthwatch (DeepSouthWatch 2019). The ontology will be used in the ENER and for the event descriptions.

5.1 Data Streaming Gathering

This application collects raw data provided by social media streaming in real-time by using the open-source Twitter streaming API (Twitter Developer 2018) to support the gathering of

information, generally represented as text. For example:

"เกิดเหตุยิงนายมุตอาลี ดีสะอะะ ซึ่งเสียชีวิต ในพื้นที่บ้านฆอลอกาเอะ จ.ยะลา"

(In English) Mr. Mutaalee Deesae was killed by shooting in Koloka-Ae village, Yala province.

This message was produced by @TichilaThaipbs at 7:47 am on May 2017. Data of this type will be used in later sections to describe each sub-process of the automatic semantic description generation. The first stage is to apply tokenization and noise removal.

5.2 Tokenization and Noise Removal

Thai is written without spaces between words, end punctuation, and capitalizations. To deal with those characteristics, this component uses a longest matching algorithm (Haruechaiyasak et al. 2008, Chaonithi and Prom-On 2016) to tokenize a line by comparing the words with dictionaries or vocabulary corpuses.

For our emergency case study, the dictionaries were automatically generated by using a taxonomy from our triple store index, and related terms from the DeepSouthwatch database (Nisalma et al. 2019) which contains 75,226 significant words. An example of the output from the tokenization component is shown at the top of Figure 4. In addition to splitting the string into tokens, special symbols and stop words are also removed, to leave only related tokens for the next stage to process effectively.

5.3 Emergency Named Entity Recognition

Named Entity Recognition (NER) is a fundamental approach in machine understanding for automatically converting tokens into knowledge. We use ENER to identify the meaning of emergency-related entities, and so create a semantic description of the event situation.

As an example, suppose that "*Yala sub-district*", "*Yala district*", and "*Yala province*" are the token entities representing locations. Therefore, "*Yala*" is the property name of the three location types, and so keyword searching will return "*Yala*" as the property name of those three resources with different types. The output will be one true positive and two false positives, an example of high recall but low precision.

In our approach, the triple store index and RexEu help deal with the above-mentioned problem. As mentioned in section 4, three key technologies are required: (1) semantic methods, (2) indexing and searching, and (3) pre-defined patterns.

Semantic technologies build the emergency ontology as a knowledge base. As detailed in section 4.1, we choose Protégé to provide an environment for developing the ontology in an OWL file for the automatic semantic descriptions. RDFlib, a Python library was employed to parse instances, classes, properties, and constraints from the OWL file in order to form the triple store. The resulting store contains 14,423 statements.

For indexing and searching, the Whoose Python library was used to order and index those statements, or TSIDs, into triples of the form (domain, predicate, and range). A graph-based structure was created with three fields - the first field stores a node as a domain, the second stores a node as a range, and the third stores the relation between those nodes. This technology was also employed for triple store searching, based on triple store index searching from section 4.3. This will bridge the rele-

vant concepts with the searching entity by using the semantic pattern in the TSIDs.

The pre-defined pattern uses regular expressions to model patterns of emergency word order such as "*Person*", "*Related object*", "*Location*", and "*Event*". RexEu utilizes 46 patterns, some of which were shown in section 4.3. Each pattern was designed by observing the characteristics of SBD in a real-world environment related to each class of emergency ontology.

The complete ENER process is shown in Figure 4.

According to Figure 4 when "*Yala province*" is part of the retrieved emergency SBD, then RexEu, acting as a process for semantic-based language dependency, will recognize it as a province entity. The "*Province*" is stored in a triple store index, and be used as its context. According to Table 1, "*Yala*" can be identified as the property name of the resource "*#Province_instance*", in the relation "*type*" with "*#Province*", and Triple index searching will return that meaning based on its context. The result is that "*Yala*" will be named without any false positives, creating a high precision, accurate emergency description.

These explicit entities, named using semantic-based language dependence, are forwarded to the emergency event description. The aim is to discover any implicit meanings, to complete the semantics of the SBD content.

5.4 Emergency Event Description and Information Creation

Emergency event description extracts the relations between named entities. These describe the SBD by defining meaning using a semantic-based bridge based on ontology constraints such as those detailed in Table 1. The extracted relations between named entities make the created emergency information more meaningful. In our case study, emergency event descriptions may use words such as "*Bombing*", "*Sabotage*", "*Arson*", and "*Shooting*". If "*Shooting*" as an "*Event*" and "*Yala*" as an instance of

Figure 4 ENER based on Triple Index and RexEu

"Province" class are found together then according to TSID number 5 from Table 1, the triple store index can identify a relationship between "Shooting" and "Yala" by creating the triple relation "#Shooting" (Subject), "#happenIn" (Predicate), and "Yala" (Object). In this way, the machine can understand and define the semantics of raw textual data. The complete example is shown in Figure 5.

The named entities and their relationships in Figure 5 are recognized in the last step, and represented graphically using the domain ontology as triple-based RDF. This event contains explicit information about the affected person, the effect, location, and the time given by eyewitnesses. This time could easily be matched against the Thai holiday calendar to obtain more meaning but that is outside the scope of the current research. The red-dashed nodes and edges show implicit entities. The red-dashed nodes in this example are #Banangstar, an instance of "District", and #Bacho, an instance of "Sub-District". The red-dashed edges are #hasDistrict, a relation between "Province" and "District", #hasSubDistrict a relation between "District" and "SubDistrict", and #hasVillage, a relation between "SubDistrict" and "Village". They represent the hidden knowledge uncovered using the semantic-based bridge build on top of the triple store index and the properties of domain and range constraints. Even though ENER only holds the locations in terms of the instances of #Province and #Village, it can travel the graph and extract the hidden relations eventually. In this

way, the triple store index allows a machine to interpret and bridge information intelligently.

The emergency information must be represented in a machine-readable format that can be used in the decision-making process. We chose RDF because its information representation matches the above graphical information. In particular, RDFlib provides a package to generate graphical information based on the RDF format, as shown in Figure 6.

The emergency event description in RDF is shown in Figure 6. RDF makes it easier to reason with the information and discover new and deeper emergency knowledge because RDF supports an intelligent system for processing and understanding information automatically. Machine can use this information during its initial phase to infer and generate emergency knowledge without human intervention. This will also help the emergency authorities make better decisions, which can reduce risk, and save human life and fundamental infrastructure.

6. Experiment Setup

The effectiveness of SBD analysis using automatic semantic description depends on the accuracy of the named entity process. We have proposed a triple store index for extending the semantic description, so our system evaluation must focus on the effectiveness of the ENER and its emergency event description components.

Figure 5 Emergency Event Description Output in Graphical Form

6.1 Experimental Objective

We wish to measure the capability of the triple store index for hidden meaning identification for named entities. The capabilities of ENER in an SBD based semantic context has two perspectives. Firstly, the semantic context is measured to interpret the word orders using semantic-based language dependence, which is important for naming the emergency-exact entities. Secondly, the semantic context using a semantic-based bridge is analyzed to make a hidden-meaning connection between emergency-exact entities.

6.2 Evaluation Metrics

This study employs three retrieval effectiveness metrics - precision, recall, and F-measure to evaluate the accuracy of the ENER process. These metrics can be calculated as $precision = \frac{TP}{TP+FP}$, $recall = \frac{TP}{TP+FN}$, and

$F - measure = \frac{2 \times (Precision \times Recall)}{Precision + Recall}$ based on three possible outcomes (TP, FP and FN). True Positive (TP) is a correct outcome of a named entity that was identified, False Positive (FP) is an incorrect outcome of a named entity that was identified, and False Negative (FN) is the related outcome of a named entity that was not identified. These metrics will be used to compare the ENER result obtained with three differed approaches - word matching ENER, semantic-based language dependent ENER, and our approach with a semantic-based bridge.

6.3 Data Set

We collected Twitter news messages concerning southern Thailand as the test SBD, between May 16 and June 16, 2018, storing 25,000 from nonprofessional sources, and 25,000 from professional journalists. However, the transac-


```

<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF
  xmlns:emergency="http://localhost/phpmyadmin#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  <emergency:province rdf:about="http://localhost/phpmyadmin#Yala">
    <emergency:hasDistrict>
      <emergency:district rdf:about="http://localhost/phpmyadmin#Banangstar">
        <emergency:districtNameTH>บันนังสตา</emergency:districtNameTH>
        <emergency:hasSubDistrict rdf:resource="http://localhost/phpmyadmin#Bacho"/>
        <emergency:districtLink rdf:resource="https://th.wikipedia.org/wiki/อำเภอบันนังสตา"/>
      </emergency:district>
    </emergency:hasDistrict>
    <emergency:provinceNameTH>ยะลา</emergency:provinceNameTH>
    <emergency:provinceLink rdf:resource="https://th.wikipedia.org/wiki/จังหวัดยะลา"/>
  </emergency:province>
  <emergency:event rdf:about="http://localhost/phpmyadmin#Shooting">
    <emergency:eventName>เกิดเหตุยิง</emergency:eventName>
    <emergency:hasAffectedPerson>
      <emergency:citizen rdf:about="http://localhost/phpmyadmin#p01">
        <emergency:personFirstName>มูดาอาลี</emergency:personFirstName>
        <emergency:personLastName>ดีสนะ</emergency:personLastName>
        <emergency:hasEffect rdf:resource="http://localhost/phpmyadmin#death"/>
        <emergency:personTitle>นาย</emergency:personTitle>
      </emergency:citizen>
    </emergency:hasAffectedPerson>
    <emergency:happenAt>
      <emergency:timestamp rdf:about="http://localhost/phpmyadmin#11052561">
        <emergency:date>11 พ.ค. 2561</emergency:date>
        <emergency:timeDescription>Morning</emergency:timeDescription>
        <emergency:time>07:47</emergency:time>
      </emergency:timestamp>
    </emergency:happenAt>
  </emergency:event>
  <emergency:sub_district rdf:about="http://localhost/phpmyadmin#Bacho">
    <emergency:hasVillage>
      <emergency:village rdf:about="http://localhost/phpmyadmin#Kholorka-e">
        <emergency:villageNo>ม.5</emergency:villageNo>
        <emergency:villageNameTH>พลองกราก</emergency:villageNameTH>
      </emergency:village>
    </emergency:hasVillage>
    <emergency:districtNameTH>บาโจ</emergency:districtNameTH>
  </emergency:sub_district>
  <emergency:effect rdf:about="http://localhost/phpmyadmin#death">
    <emergency:effectName>เสียชีวิต</emergency:effectName>
  </emergency:effect>
</rdf:RDF>

```

Figure 6 Emergency Event Description Output in RDF Format

tions from nonprofessional journalists were often duplicated, rumors, or fakes so we manually selected only factual transactions which reduced the number to 18,504. Professional journalists accounted for 10,553 of the transactions, with the rest coming from nonprofessional sources. All the messages were in Thai.

6.4 Results and Discussions

Three ENER approaches: word-based matching (Zhao et al. 2016, Rossi et al. 2018), semantic-based language dependence (Jiménez-Zafra et al. 2018, Piao et al. 2019), and our semantic-based bridge were judged

in terms of average precision, recall, and F-measure. Average precision measures how well the ENER can identify relevant entities, and eliminate irrelevant entities. Average recall measures how well ENER can successfully identify relevant entities. Average F-measure (also called the harmonic mean) averages precision and recall, to show the robustness of those approaches.

We categorized ENER outcomes into five classes taken from the proposed ontology given in section 4.1. For instance, the class "Person" has sub-classes "Citizens", "Authorities", and "Criminals". We considered the results for

each class separately because each class has its own ENER characteristics. For example, the "Location" name in Thailand may be used for multiple types of location, and a single location can have several names. However, we did not consider the "Time" class because it was delivered from time-stamp information that is already structured represented in a machine-readable form.

The effectiveness results for the three approaches are shown in Table 2. The average F-measure comparison of the approaches is given in Figure 7.

Table 2 compares three ENER approaches: word-based matching, semantic-based language dependency, and our semantic-based bridge, based on precision, recall, and F-measure.

In word-based matching, the overall averages are quite low, but the entity "Criminals", "Provinces", "Emergency Events", "Related Objects" show high accuracy compared to the other approaches. This is because those entities are distinctively and explicitly represented by a single word in SBD. This makes this approach suitable for systems that only need to match and define the meaning of exact key features (Zhao et al. 2016, Rossi et al. 2018). The strong point of this approach is that it utilizes a simple algorithm using a dictionary-based process, and directly matches SBD entities and featured keys from the dictionary that define the meanings of the entities. If the system only needs some information about "Criminals", "Provinces", and "Emergency Events", then this approach is good enough. However, the quality of this approach depends on the completeness of the dictionary space. It cannot solve complex problems, especially in emergencies where word orders have low overall accuracy.

The overall accuracy for the semantic-based language dependency approach is higher than for word-based matching. This approach is more fitting for SBD entities that can be represented in natural language and captured with

regular patterns. It can deal very well with the dependent word order problem. Research has shown that this approach is acceptable for recognizing and naming emergency entities (Jiménez-Zafra et al. 2018, Piao et al. 2019). However, unlike word-based matching, this approach needs experts to design the semantic-based language dependency patterns. Moreover, it cannot deal with hidden entities.

Our semantic-based bridge approach deals with the characteristics of the SBD problem, and its overall accuracy is higher than the other approaches because it can discover hidden meanings needed for emergency management. Although the semantic-based language dependency approach is satisfactory, emergency events need higher accuracy that can support decision making. For example, the accuracies of the classes "Public places", "Sub-Districts" and "Districts", "Authorities" and "Citizens" are much higher than in the other approaches since those entities are represented by hidden meanings which our approach exposes.

The comparative effectiveness of the three approaches using F-measures is shown in Figure 7, as a summary of our findings.

Figure 7 represents a comparison of the average F-measures as a histogram, with our semantic-based bridge approach producing the highest values, indicating that it has the highest accuracy and robustness for recognizing emergency entities from SBD. This is due to its ability to deal with entities represented indirectly. Ambiguous meaning is a major problem for informal and natural language, and can definitely be found in SBD.

These results show the effectiveness of our ENER approach using a semantic-based bridge with a triple store index since its named entities can be used to support automatic semantic descriptions for accurate emergency information. This is especially true for information about incident locations, event types, and victim details, which are fundamental requirements for emergency management. This shows that our

Table 2 The Comparative Effectiveness of Three ENER Approaches

Classes	Entities	Word-based matching			Semantic-based language dependency			Semantic-based bridge (Our proposal)		
		Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Person	Citizens	0.27	0.50	0.35	0.81	0.68	0.74	0.86	0.95	0.90
	Authorities	0.54	0.40	0.46	0.86	0.71	0.78	0.93	0.96	0.95
	Criminals	0.82	0.62	0.71	0.93	0.97	0.95	1.00	0.97	0.98
Location	Provinces	0.71	0.87	0.78	0.91	0.87	0.89	0.99	0.97	0.98
	Sub-Districts	0.39	0.54	0.45	0.61	0.70	0.65	0.92	0.96	0.94
	Districts	0.25	0.33	0.29	0.51	0.65	0.57	0.86	0.86	0.86
	Villages	0.57	0.66	0.61	0.76	0.89	0.82	0.82	0.84	0.83
	Public places	0.43	0.63	0.51	0.73	0.78	0.76	0.98	0.98	0.98
	Government places	0.32	0.57	0.41	0.72	0.93	0.81	1.00	0.81	0.90
Event	Emergency events	0.88	0.81	0.84	0.99	0.81	0.89	0.99	0.93	0.96
Object	Related objects	0.73	0.85	0.79	0.97	0.85	0.90	1.00	0.97	0.98
Effect	Lifeless effects	0.56	0.45	0.50	0.91	0.91	0.91	1.00	0.91	0.95
	Injured effects	0.91	0.50	0.65	0.89	0.80	0.84	1.00	0.95	0.97
Overall average		0.57	0.59	0.56	0.82	0.81	0.81	0.95	0.93	0.94

Figure 7 The Average F-measure Comparison between the Three Approaches

proposed semantic description can help authorities make better decisions.

However, some entities still have the challenges to improve their accuracies such as "Government places", "Districts", "Villages", and "Citizens" could be approved since their precision or recall is lower than 90%. One limitation that needs to be addressed is the use of the relative pronoun by non-professional journalists in their messages.

7. Conclusions

This paper proposes a Social Big Data (SBD) analysis using automatic semantic descriptions to support an emergency management system. Our main contribution is to apply the triple index approach to Emergency Named Entity Recognition (ENER) to handle both explicit and implicit emergency information generated from natural language in the SBD. The aim is to increase average precision, to enhance the

accuracy of the semantic emergency information. Our ENER approach is a combination of word matching and semantic analysis based on word orders for explicit information, extended with a semantic-based bridge to deal with implicit information. The results show that our approach gives higher average accuracy than word matching and language-based dependencies. The improvement is especially useful in the field of emergency management where the safety of citizens and the stability of government infrastructure depend on accurate information for fast and precise decision-making.

Our approach is focused on automatic semantic descriptions based on single transactions. However, SBD usually emerges from multiple transactions. The relationship and meaning between these transactions must be considered to better describe events. The ability to understand an emergency event affected by the transaction movement would reveal deeper trends in the situation. Moreover, our approach focuses on a single datatype and a relatively small sample, which may lead to overfitting. In future work, we plan to analyze emergency information from SBD by considering the relationships among transactions. We will also expand our emergency knowledge-base and triple index matching algorithms to address natural language problems such as wrong spelling and local names written in different ways. Two main issues are how to deal with uncertainty in the SBD characteristics by utilizing machine learning, and how to handle unreliable crowd sourcing accounts.

Acknowledgments

The authors would like to thank the anonymous reviewers for their time and effort. Their constructive comments and helpful suggestions helped us to clarify the main paper's research contributions and improve its quality.

References

- Bello-Organ G, Jung JJ, Camacho D (2016). Social big data: Recent achievements and new challenges. *Information Fusion* 28:45-59. <https://doi.org/10.1016/j.inffus.2015.08.005>.
- Bitencourt K, Araújo Durão F, Mendonça M, Laique Bomfim de Souza Santana L (2018). An ontological model for fire emergency situations. *IEICE Transactions on Information and Systems*. E101D:108-115.
- Castillo C (2016). *Big Crisis Data: Social Media in Disasters and Time-Critical Situations*. Cambridge University Press, New York, NY.
- Chaonithi K, Prom-On S (2016). A hybrid approach for Thai word segmentation with crowdsourcing feedback system. *2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, ECTI-CON 2016*: 1-6.
- Choi S, Bae B (2015). The real-time monitoring system of social big data for disaster management. *Lecture Notes in Electrical Engineering* 330:809-815.
- DeepSouthWatch.org | Deep SouthWatch Database (2019). DeepSouthWatch.
- Derczynski L, Maynard D, Rizzo G (2015). Analysis of named entity recognition and linking for tweets. *Information Processing and Management* 51:32-49.
- Developer.twitter.com (2018). Twitter Developer.
- Farag MMG, Lee S, Fox EA (2018). Focused crawler for events. *International Journal on Digital Libraries* 19:3-19.
- Fernández N, Fuentes D, Sánchez L, Fisteus JA (2010). The NEWS ontology: Design and applications. *Expert Systems with Applications* 37:8694-8704.
- Guo K, Tang Y, Zhang P (2017). CSF: Crowdsourcing semantic fusion for heterogeneous media big data in the internet of things. *Information Fusion* 37:77-85.
- Haruechaiyasak C, Kongyoung S, Dailey M (2008). A comparative study on thai word segmentation approaches. *5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, ECTI-CON 2008* 1:125-128.
- Jiménez-Zafra SM, Taulé M, Martín-Valdivia MT (2018). SFU ReviewSP-NEG: A Spanish corpus annotated with negation for sentiment analysis. A typology of negation patterns. *Language Resources and Evaluation* 52:533-569.
- Jing M, Scotney BW, Coleman SA (2016). Flood event image recognition via social media image and text analysis. In: *Signals and Systems Conference (ISSC)*, 2016 27th Irish:4-9.
- Kuflik T, Minkov E, Nocera S (2017). Automating a framework to extract and analyse transport related social media content: The potential and the challenges. *Transportation Research Part C: Emerging Technologies* 77:275-291.
- Li S, Chen S, Liu Y (2019). A method of emergent event evolution reasoning based on Ontology Cluster and Bayesian Network. *IEEE Access* 7:15230-15238.
- Meng X, Xu C, Liu X (2018). An Ontology-Underpinned emergency response system for water

- pollution accidents. *Sustainability*. <https://doi.org/10.3390/su10020546>.
- Murphy RR (2016). Emergency informatics: Using computing to improve disaster management. *Computer* 20:19-27.
- Murthy D, Gross AJ (2017). Social media processes in disasters: Implications of emergent technology use. *Social Science Research* 63:356-370.
- Nisalma Kuseng, Sulinda Muangsuk, Wanee Desulong (2019). Deep South Watch Database. DeepSouthWatch.
- Olshannikova E, Olsson T, Huhtamäki J, Kärkkäinen H (2017). Conceptualizing big social data. *Journal of Big Data*. <https://doi.org/10.1186/s40537-017-0063-x>.
- Oramas S, Espinosa-Anke L, Sordo M (2016) Information extraction for knowledge base construction in the music domain. *Data and Knowledge Engineering* 106:70-83.
- Pham P, Nguyen T, Do P (2018). Computing domain ontology knowledge representation and reasoning on graph database. In: *Information Systems Design and Intelligent Applications*. Springer Singapore.
- Piao S, Rayson P, Knight D, Watkins G (2019). Towards a Welsh semantic annotation system. In: *LREC 2018 - 11th International Conference on Language Resources and Evaluation*: 980-985.
- Poblet M, García-Cuesta E, Casanovas P (2018). Crowdsourcing roles, methods and tools for data-intensive disaster management. *Information Systems Frontiers* 20:1363-1379.
- Riccardi MT (2016). The power of crowdsourcing in disaster response operations. *International Journal of Disaster Risk Reduction* 20:123-128.
- Rossi C, Acerbo FS, Ylinen K, et al (2018). Early detection and information extraction for weather-induced floods using social media streams. *International Journal of Disaster Risk Reduction* 30:145-157.
- Sahoh B, Choksuriwong A (2018). Towards smart emergency management: Trends and challenges of feature engineering. In: *2018 22nd International Computer Science and Engineering Conference (ICSEC)*. IEEE:1-4.
- Sahoh B, Choksuriwong A (2017). Smart emergency management based on social big data analytics: Research trends and future directions. In: *Proceedings of the 2017 International Conference on Information Technology - ICIT 2017*. 1-6.
- Sandhu R, Gill HK, Sood SK (2016). Smart monitoring and controlling of Pandemic Influenza A (H1N1) using Social Network Analysis and cloud computing. *Journal of Computational Science* 12:11-22.
- Sayed A, Al Muqrishi A (2017). IBRI-CASANTO: Ontology-based semantic search engine. *Egyptian Informatics Journal* 18:181-192.
- Shah R, Zimmermann R (2017). *Multimodal Analysis of User-Generated Multimedia Content*. Springer International Publishing.
- Stepanov EA, Chowdhury SA, Bayer AO (2018). Cross-language transfer of semantic annotation via targeted crowdsourcing: Task design and evaluation. *Language Resources and Evaluation* 52:341-364.
- Stork L, Weber A, Gassó Miracle E (2018). Semantic annotation of natural history collections. *Journal of Web Semantics*. <https://doi.org/10.1016/j.websem.2018.06.002>.
- Velasco-Elizondo P, Marín-Piña R, Vazquez-Reyes S (2016). Knowledge representation and information extraction for analysing architectural patterns. *Science of Computer Programming* 121:176-189.
- Xu Z, Liu Y, Yen N (2016). Crowdsourcing based description of urban emergency events using social media big data. *IEEE Transactions on Cloud Computing*: 1-11.
- Yang B, Wang J, Ma W, Mao X (2018). Research on crowdsourcing emergency information extraction of based on event's frame. *Journal of Physics: Conference Series*. <https://doi.org/10.1088/1742-6596/933/1/012009>.
- Zhao Y, Zhang J, Yao L (2016). Object-oriented information extraction and evaluation of seismic damage of buildings using very high spatial resolution imagery. *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*: 2853-2856.
- Zhou P, El-Gohary N (2017) Ontology-based automated information extraction from building energy conservation codes. *Automation in Construction* 74:103-117.
- Bukhoree Sahoh** received his B.S. degree in computer science from Prince of Songkla University (PSU), Songkla, Thailand, in 2011, and a MSc degree in information technology from Thaksin University, Phatthalung, Thailand, in 2015. He is currently pursuing a Ph.D. degree in computer engineering at PSU, and is a research assistant in PSU's Intelligent Automation Research Center. His research interests include natural language processing, complex event understanding, explainable artificial intelligence, and causality in machine learning.
- Anant Choksuriwong** is a lecturer in the Department of Computer Engineering, Prince of Songkla University (PSU), Songkla, Thailand, and a member of the committee of the Intelligent Automation Research Center at PSU. He received his bachelor's degree in computer engineering from PSU in 2000, and master's degrees from Université Joseph Fourier, and the Institut National Polytechnique de Grenoble in Imagerie, Vision, and Robotique, France, in 2003 and 2004. Anant Choksuriwong earned his Ph.D. in Sciences et Technologies industrielles at Université d'Orléans, France, in 2008. His current research emphasizes highly adaptive artificial intelligence, machine learning, and cognitive systems engineering.